1. Resumen

El avance tecnológico y la sofisticación de las computadoras ha dado lugar a procesadores capaces de ejecutar millones de operaciones por segundo sobre grandes volúmenes de datos. Paralelamente a este avance, ha aparecido en los últimos años el concepto de *Data Mining* o la *Minería de Datos*, que es una rama de las ciencias de la computación cuya finalidad última es la elaboración de patrones y la extracción de información de grandes bases de datos.

El objetivo principal del presente proyecto es el análisis de los datos académicos de los estudiantes de Grado en Ingeniería en Tecnologías Industriales de L'Escola Tècnica Superior d'Enginyeria Industrial de Barcelona (ETSEIB). La finalidad última de dicho análisis es la construcción de distintos modelos que permitan predecir el comportamiento del estudiante a lo largo de los distintos cursos. En dicho proyecto, el término comportamiento hace referencia a una serie de aspectos tales como si el alumno repetirá alguna asignatura, si estará por encima o por debajo de la media general del curso, o si aprobará los cursos al ritmo estipulado (curso por año). Además de los modelos de comportamiento, también se desea construir distintas expresiones que permitan estimar la nota que obtendrá el alumno en cada una de las asignaturas en función de las notas obtenidas en cursos anteriores.

Para lograr dicho objetivo, se han aplicado técnicas estadísticas muy variopintas en función de la finalidad deseada y la naturaleza de los datos a tratar. Para el estudio de significación de variables, se ha recurrido al uso de técnicas no paramétricas como la prueba de la U de Mann Whitney o el test de Kruskal Wallis para las variables cuantitativas, o la prueba de Chi - Cuadrado para las variables categóricas. Para la construcción de los modelos de comportamiento, la técnica utilizada ha sido la regresión logística binaria, que es la que ha ofrecido unos resultados más precisos. Por último, la estimación de la nota de cada una de las asignaturas se ha realizado mediante una combinación lineal de las notas de asignaturas ya cursadas, empleando la técnica de la regresión lineal.

La precisión de cada uno de los modelos generados se ha medido con un subconjunto de datos que no han participado en la creación del mismo. Los porcentajes globales de acierto obtenidos para los modelos de comportamiento oscilan entre el 75,0 y el 83,0 %, mientras que las cualificaciones que obtendrá el alumno en cada una de las asignaturas se pronostica con un 60,1 % de acierto en promedio. Si un alumno repetirá o no una asignatura en concreto, se pronostica con un 82,4 % de acierto en promedio. Estos porcentajes de acierto permitirán realizar un análisis de los flujos de estudiantes más preciso y optimizar los recursos, tanto de personal, como de material, como de espacio. Además, también se podrá ofrecer una mayor ayuda a aquellos estudiantes con mayor probabilidad de fracaso.

2. Sumario

3. Introducción

La estadística es una ciencia cuyos inicios se encuentran en el siglo XVIII. En primera instancia, se utilizaba el término estadística para hacer referencia a la colección sistemática de datos económicos y demográficos. Con el paso del tiempo y el avance del conocimiento, la palabra estadística ha ampliado su significado, incluyendo además de la recolección de datos, el análisis de los mismos para la búsqueda de explicaciones, dependencias, y conclusiones sobre un determinado fenómeno. En la actualidad, es muy empleado el término *Data Mining* (o Minería de Datos) para hacer referencia a la computación de datos a gran escala con el objetivo de descubrir patrones, utilizando métodos como la inteligencia artificial, el aprendizaje automático, y la estadística.

Las áreas de aplicación tanto del *Data Mining* como de las distintas técnicas estadísticas son muy diversas, y los estudios realizados pueden ir desde el ámbito social o sanitario, hasta el área de los negocios o los controles de calidad.

Con el presente proyecto se pretende realizar un análisis estadístico de los estudiantes de grado de la titulación Grado en Ingeniería en Tecnologías Industriales impartida en la Escola Tècnica Superior d'Enginyeria Industrial de Barcelona (ETSEIB). La finalidad de dicho análisis es múltiple. En primer lugar, se desea realizar una presentación de la titulación, mostrando cómo está estructurada, presentando aquellas estadísticas descriptivas más relevantes y realizando un análisis en profundidad de aquellas asignaturas con resultados más extremos. En segundo lugar, se realiza una estimación de los resultados obtenidos en cada una de las asignaturas del primer curso en función de las notas obtenidas en la selectividad. En tercer lugar, se desea construir un modelo capaz de predecir el comportamiento de los estudiantes en cada uno de los cursos basándose en una serie de parámetros calculados para cada uno de los estudiantes. En el caso que ocupa este proyecto, el término comportamiento engloba una serie de variables que definen al estudiante, como por ejemplo si éste aprueba los cursos al ritmo marcado (curso por año), si tiene una media superior o inferior a media total de cada curso, o si el alumno repetirá o no en los distintos cursos. Por último, se desea realizar una estimación de las notas obtenidas en cada una de las asignaturas mediante una combinación lineal de las asignaturas cursadas anteriormente. Debido a que durante el primer cuatrimestre de primero aún no se dispone de datos de asignaturas anteriores, dicha estimación se realizará para todas las asignaturas del segundo cuatrimestre en adelante.

Para llevar a cabo el análisis y la posterior construcción de los modelos, es necesario comprobar si existen diferencias estadísticamente significativas en variables como el sexo y la ubicación del estudiante. Para ello, se llevan a cabo distintos contrastes estadísticos, que

son seleccionados en función de la naturaleza de las variables a analizar (cuantitativas o categóricas). Con la aplicación de dichos contrastes estadísticos sobre las variables, además de la información aportada, se pretende conocer si dichas variables deben ser consideradas o no a la hora de construir los modelos de comportamiento.

Finalmente, es necesario validar todos los modelos creados para evaluar su precisión. Para ello, resulta especialmente importante crear un subconjunto de validación que no participe en la creación del modelo. Evaluar la capacidad del modelo con datos que han sido incluidos en la creación del mismo puede llevar a la obtención de unos resultados más precisos a los que se obtendrían en la vida real.

3.1. Objetivo del proyecto

El objetivo del proyecto es la aplicación de distintas técnicas estadísticas con el programa SPSS con el fin de realizar un análisis sobre los datos académicos de los estudiantes de Grado en Ingeniería en Tecnologías Industriales de l'*Escola Tècnica Superior d'Enginyeria Industrial de Barcelona* (ETSEIB). La finalidad última de dicho análisis es proporcionar una descripción numérica de la titulación y generar distintos modelos que permitan determinar aspectos como si el alumno repetirá o no, si estará por encima o por debajo de la media, o si completará la titulación yendo a curso por año. Además, se desea estimar también las notas que sacará el estudiante en cada una de las distintas asignaturas, permitiendo predecir cuáles suspenderá y en cuales obtendrá una nota igual o superior al 5. Con todo ello, se pretende que el profesorado pueda anticiparse a los distintos sucesos, pudiendo así optimizar los recursos, ya sean de espacio, materiales, o de personal.

3.2. Alcance del proyecto

Existen distintas técnicas a la hora de abordar las predicciones realizadas sobre los estudiantes, como por ejemplo la inteligencia artificial o el aprendizaje automático. No obstante, el presente proyecto se ha centrado únicamente en la aplicación de técnicas estadísticas con el software SPSS. Dicho software cuenta con una interface amigable para el tratamiento de datos parecida a la de Microsoft Excel, y tiene integrados distintos paquetes estadísticos para aplicar las pruebas necesarias sin necesidad de programar. Pese a que dicho programa ofrece una gran cantidad de técnicas estadísticas, no todas han sido contempladas en dicho proyecto para limitar la extensión de este. Como máximo, se han valorado dos técnicas para la construcción de los modelos, que son la regresión logística binaria y el análisis discriminante.

4. Datos de origen

En este capítulo se presentarán las tres bases de datos de las que se dispone para el presente proyecto y se mostrarán las transformaciones necesarias para realizar el estudio.

La información de cada una de las bases de datos es la siguiente: en la primera dan los datos referentes a la fase inicial de la universidad (primer curso); en la segunda se muestran los datos de la fase no inicial (cursos posteriores a primero); y la última de ellas contiene los datos personales de cada uno de los alumnos.

En primer lugar, se realizará una breve descripción a nivel general de cada una de las tres bases de datos, mostrando el número de registros y los campos que tiene cada uno de ellos. En los casos que sea preciso, se dará una breve explicación de los datos contenidos en dichos campos. En segundo lugar, se explicará qué registros han sido eliminados y por qué, y por último, se presentarán las nuevas variables creadas con el objetivo de sintetizar información y de colocarla en una disposición que facilite el estudio.

4.1. BBDD de Datos personales de los alumnos

Es la base de datos menos extensa de las tres. Cuenta con 2951 registros referentes a 2781 alumnos. El hecho de que haya más registros que alumnos se debe a distintos motivos, como por ejemplo alumnos que se han cambiado de titulación, alumnos que han cursado varias titulaciones, o alumnos que han empezado el plan de estudios antiguo y se han adaptado al Plan Bolonia.

Los campos que contiene dicha base de datos son los siguientes:

- ID Alumno: es un número identificativo de seis cifras único para cada alumno.
- Sexo.
- Código Postal del lugar de residencia.
- Año de ingreso en la Universidad Politécnica de Catalunya.
- Vía de acceso a la universidad.
- Nota de la selectividad.
- Código de le escuela de procedencia del alumno.

- Código Postal de la escuela de procedencia.

Los años de ingreso en la universidad de los alumnos van desde 1987, que corresponde al registro con el año de ingreso más antiguo, hasta 2013 que son los alumnos de los que se posee datos que han ingresado más recientemente.

Por lo que se refiere a las vías, se pueden observar hasta doce vías de acceso diferentes para los alumnos, dependiendo de si han accedido realizando la selectividad, una formación profesional, un cambio de titulación, o cualquier otro modo de acceso a la universidad.

Tal y como se verá más adelante, en este proyecto se realizará un filtrado de estudiantes en función de sus características, ya que es especialmente importante descartar todos aquellos datos que por su frecuencia se pueden considerar excepcionales, ya que pueden distorsionar el estudio realizado y dar lugar a conclusiones erróneas.

4.2. BBDD de la Fase inicial y de la Fase no inicial

Tal y como se ha mencionado en la primera parte de este capítulo, los datos de la fase inicial (primer curso) y de la fase no inicial (cursos posteriores al primero) vienen dados en bases de datos distintas. La base de datos de la fase inicial cuenta con 33.385 registros de 3.045 alumnos distintos, mientras que la de la fase no inicial está formada por 38.922 registros correspondientes a 1.831 alumnos distintos. A continuación se resumen las características principales de ambas bases de datos:

	Fase Inicial	Fase No Inicial
Nº Registros	33.385	38.922
Nº Alumnos	3.045	1.831
Nº Asignaturas distintas	12	156
Años de Inicio	2010 - 2014	2010 - 2014
Titulaciones	648, 752, 753 754, 823, 864	752, 753, 754

Tabla 4.1. Características principales de los datos en las BBDD

Los campos de las dos bases de datos son los mismos, y son los siguientes:

- ID Alumno

- Titulación cursada
- Asignatura realizada
- Número de créditos de la asignatura
- Año de inicio de la asignatura
- Cuatrimestre (Otoño o Primavera)
- Asignatura aprobada o suspendida
- Nota puesta por el profesor
- Nota curricular
- Nota final de la asignatura
- Grupo en el que se ha realizado la asignatura

Dado que se aplicarán distintas transformaciones a la base de datos original, y una de ellas es añadir un campo nuevo con el curso al que pertenecen las asignaturas, la base de datos de la fase inicial y la de la fase inicial se trataran de manera conjunta, ya que los campos para éstas dos son los mismos. De este modo, se obtiene una única base de datos de 72.307 registros.

En este caso, también será necesario realizar un filtrado de la base de datos para centrarse en aquellos alumnos objeto de estudio de este proyecto. Con este fin, es necesario eliminar todos aquellos datos que puedan interferir de manera negativa en el modelo, o que simplemente no aporten nada.

4.3. Estudiantes objeto de estudio

Debido a la gran diversidad de los datos disponibles, se debe acotar la base de datos y abordar únicamente aquellos estudiantes que responden a unas características determinadas. El estudio realizado en este proyecto se centra en aquellos estudiantes que han entrado en la Universidad Politécnica de Cataluña estando ya en vigor el Plan Bolonia, que han accedido por la vía de la selectividad y que cursa la titulación de Grado en Ingeniería en Tecnologías Industriales que se imparte en la ETSEIB. Además de la titulación en la que se centrará el estudio, en la ETSEIB se imparten dos titulaciones más de Grado:

- **753:** Grado en Ingeniería Química
- **754:** Grado en Ingeniería de Materiales

Pese a que se mostrará algún dato general de estas dos titulaciones, no se estudiarán en profundidad debido al insuficiente número de datos, ya que el 80 % de los registros corresponden a estudiantes de Grado en Ingeniería en Tecnologías Industriales.

Con el fin de quedarse únicamente con estos estudiantes, se eliminan los siguientes registros de la base de datos:

- Aquellos registros correspondientes a alumnos que cursan las titulaciones 864, 648 y 823 (se dejan los de las titulaciones 753 y 754 para poder extraer algún dato general)
- Los registros correspondientes a 388 alumnos sin sus datos generales.
- Los datos correspondientes a 93 alumnos que han ingresado antes del 2010.
- Los registros pertenecientes a 51 alumnos que no han accedido mediante selectividad.
- Registros en que la asignatura ha sido convalidada o falta algún dato.

Además, con el objetivo de colocar la información en una disposición más óptima para realizar el análisis, los datos personales de los alumnos se colocarán en cada registro de la base de datos que contiene la fase inicial y no inicial. De este modo, la base de datos de los datos personales de los alumnos no se empleará como tal, sino que su información estará contenida en la base de datos resultante de juntar las dos fases.

Finalmente, queda una única base de datos con 65.178 registros correspondientes a 2.295 alumnos, aunque el análisis en profundidad se realizará únicamente para los 1820 alumnos de la titulación 752. El reparto es el siguiente:

Titulación	Nº Registros	Nº Alumnos
752	54993	1820
753	7050	304
754	3135	171
Total	**65178**	**2295**

Tabla 4.2. Resumen de las características de la base de datos.

Partiendo de estos datos de inicio, se han creado distintas variables y se han realizado diversas modificaciones sobre la base de datos para poder realizar un mejor análisis. Estos cambios de muestran en el siguiente apartado.

4.4. Transformaciones de la base de datos original

Como ya se ha dicho, se han realizado diversas transformaciones sobre las bases de datos originales con el objetivo de disponer los datos de mejor manera para su análisis. Recordemos brevemente que se han juntado las tres bases de datos en una única, y que se han eliminado los registros de estudiantes que:

- No cursan alguna de las tres titulaciones de grado que se hacen en la ETSEIB (aunque el estudio se centrará únicamente en una).
- Han ingresado antes de 2010 o no han accedido mediante la selectividad.
- Tienen asignaturas convalidadas.
- Tienen algún campo de sus datos personales incompleto.

Para cada uno de los registros de la base de datos se han añadido distintos campos con el objetivo de sintetizar, codificar o ampliar información. A continuación se explican las variables añadidas.

4.4.1. Creación de variables

Con la información procedente de las tres bases de datos originales, se ha creado una única con los campos que se muestran más abajo. Para la creación de algunos de ellos, ha sido necesario la programación con el módulo de Visual Basic que proporciona Microsoft Excel.

- ID Alumno: Número identificativo de seis dígitos único para cada alumno.
- SexoCod: Sexo codificado. Toma el valor 0 para las mujeres, y 1 para los hombres.
- Ubicación: Indica el lugar de residencia del alumno. Esta variable se ha codificado. Sus valores son los siguientes:
    - 1 - Barcelona
    - 2 - Tarragona
    - 3 - Lleida

  - 4 - Girona
  - 5 - Islas Baleares
  - 6 - Otros

- Código Escuela: Es un código que indica la escuela de procedencia del alumno.
- Ubicación Escuela: Indica el lugar donde se ubica la escuela. Sigue la misma codificación que la variable "Ubicación".
- CP Alumno: Código postal del lugar de residencia.
- Año Ingreso: Año en que ha ingresado en la UPC.
- Nota_Sele: Nota que el alumno ha sacado en la selectividad.
- Grupos_Sele: Se agrupan los estudiantes en función de su nota de selectividad.
  - Entre un 5,00 y 6,99 $\rightarrow$ Grupos_Sele = 1
  - Entre un 7,00 y 8,99 $\rightarrow$ Grupos_Sele = 2
  - Entre un 9,00 y 10,99 $\rightarrow$ Grupos_Sele = 3
  - Entre un 11,00 y 12,99 $\rightarrow$ Grupos_Sele = 4
  - Más de un 13 $\rightarrow$ Grupos_Sele = 5
- Creditos Susp 1ero: Número de créditos suspendidos en primero por el alumno.
- Creditos Apr 1ero: Número de créditos aprobados en primero por el alumno.
- Total Creditos 1ero: Número total de créditos cursados por el alumno en primero.
- Promedio 1ero: Es la media ponderada de las asignaturas realizadas en primero. Se calcula de la siguiente manera:

$$\mathrm{Promedio\ 1ero} = \frac{\sum_{i=1}^{n} \left(Nota_i \cdot N^o Creditos_i \right)}{\sum_{i=1}^{n} N^o Creditos_i}$$

- Grupos_Primero: Puede tomar los valores 0 y 1. Un 0 indica que ese alumno tiene una nota media en primero por debajo de la nota media general de primero, mientras que un 1 indica que ese alumno tiene una nota media en primero superior a ésta.

- Repetidas_Primero: Número de veces que se ha repetido una o varias asignaturas en primero.

- Repiten_Primero_S_N: toma el valor 0 si el alumno no ha repetido ninguna asignatura en primero, y 1 si ha repetido una o más asignaturas.

- Alpha_Primero: Parámetro relacionado con el número de créditos aprobados respecto al total en primero. Se calcula de la siguiente manera:

$$Alpha\_\mathrm{Pr}imero = \frac{Creditos\_Apr\_1ero}{Total\_Creditos\_1ero}$$

- Completan_1ero_S_N: Toma el valor 1 si el alumno ha completado primero y 0 si todavía no lo ha completado.

- Completan_1ero_Al_Ritmo: Toma el valor 1 si el alumno completa el primer curso en el primer año, y 0 si no lo completa o lo completa en más de un año.

- Creditos Susp 2ndo: Número de créditos suspendidos en segundo por el alumno.

- Creditos Apr 2ndo: Número de créditos aprobados en segundo por el alumno.

- Total Creditos 2ndo: Número total de créditos cursados por el alumno en segundo.

- Promedio 2ndo: Es la media ponderada de las asignaturas realizadas en segundo.

- Grupos_Segundo: Puede tomar los valores 0 y 1. Un 0 indica que ese alumno tiene una nota media en segundo por debajo de la nota media general de segundo, mientras que un 1 indica que ese alumno tiene una nota media en segundo superior a ésta.

- Repetidas_Segundo: Número de veces que se ha repetido una o varias asignaturas en segundo.

- Repiten_Segundo_S_N: toma el valor 0 si el alumno no ha repetido ninguna asignatura en segundo, y 1 si ha repetido una o más asignaturas.

- Alpha_Segundo: Parámetro relacionado con el número de créditos aprobados respecto al total en segundo. Se calcula de la siguiente manera:

$$Alpha\_Segundo = \frac{Creditos\_Apr\_2ndo}{Total\_Creditos\_2ndo}$$

- Completan_2ndo_S_N: Toma el valor 1 si el alumno ha completado segundo y 0 si todavía no lo ha completado.

- Completan_2ndo_Al_Ritmo: Toma el valor 1 si el alumno completa el segundo curso en los dos primeros años, y 0 si no lo completa o lo completa en más de dos años.

- Creditos Susp 3ero: Número de créditos suspendidos en tercero por el alumno.

- Creditos Apr 3ero: Número de créditos aprobados en tercero por el alumno.

- Total Creditos 3ero: Número total de créditos cursados por el alumno en tercero.

- Promedio 3ero: Es la media ponderada de las asignaturas realizadas en tercero.

- Grupos_Tercero: Puede tomar los valores 0 y 1. Un 0 indica que ese alumno tiene una nota media en tercero por debajo de la nota media general de tercero, mientras que un 1 indica que ese alumno tiene una nota media en tercero superior a ésta.

- Repetidas_Tercero: Número de veces que se ha repetido una o varias asignaturas en tercero.

- Repiten_Tercero_S_N: toma el valor 0 si el alumno no ha repetido ninguna asignatura en tercero, y 1 si ha repetido una o más asignaturas.

- Alpha_Tercero: Parámetro relacionado con el número de créditos aprobados respecto al total en tercero. Se calcula de la siguiente manera:

$$Alpha\_Tercero = \frac{Creditos\_Apr\_3ero}{Total\_Creditos\_3ero}$$

- Completan_3ero_S_N: Toma el valor 1 si el alumno ha completado tercero y 0 si todavía no lo ha completado.

- Completan_3ero_Al_Ritmo: Toma el valor 1 si el alumno completa el tercer curso en los tres primeros años, y 0 si no lo completa o lo completa en más de tres años.

- Creditos Susp 4rto: Número de créditos suspendidos en cuarto por el alumno.
- Creditos Apr 4rto: Número de créditos aprobados en cuarto por el alumno.
- Total Creditos 4rto: Número total de créditos cursados por el alumno en cuarto.
- Promedio 4rto: Es la media ponderada de las asignaturas realizadas en cuarto.
- Grupos_Cuarto: Puede tomar los valores 0 y 1. Un 0 indica que ese alumno tiene una nota media en cuarto por debajo de la nota media general de cuarto, mientras que un 1 indica que ese alumno tiene una nota media en cuarto superior a ésta.
- Repetidas_Cuarto: Número de veces que se ha repetido una o varias asignaturas en cuarto.
- Repiten_Cuarto_S_N: toma el valor 0 si el alumno no ha repetido ninguna asignatura en cuarto, y 1 si ha repetido una o más asignaturas.
- Alpha_Cuarto: Parámetro relacionado con el número de créditos aprobados respecto al total en cuarto. Se calcula de la siguiente manera:

$$Alpha\_Cuarto = \frac{Creditos\_Apr\_4rto}{Total\_Creditos\_4rto}$$

- Completan_4rto_S_N: Toma el valor 1 si el alumno ha completado cuarto y 0 si todavía no lo ha completado.
- Completan_4rto_Al_Ritmo: Toma el valor 1 si el alumno completa el cuarto curso en los cuatro años de carrera, y 0 si no lo completa o lo completa en más de cuatro años.
- Aleatorio: se asigna un número aleatorio entre 0 y 100 para poder dividir posteriormente la base de datos en subconjuntos de entrenamiento y de validación.

4.5. Titulación Grado en Ingeniería en Tecnologías Industriales

Como ya se ha dicho anteriormente, el presente proyecto se centrará en la titulación Grado en Ingeniería en Tecnologías Industriales, ya que se dispone de gran cantidad de datos para generar un modelo predictivo para cada uno de los cursos. Dicha titulación consta de 207

créditos troncales, 21 optativos, y 12 pertenecientes al Trabajo de Fin de Grado (TFG), sumando un total de 240 créditos repartidos en 4 cursos:

	Troncales	Optativos	TFG
Primero	60	0	0
Segundo	57	3	0
Tercero	60	0	0
Cuarto	30	18	12
TOTAL	**207**	**21**	**12**

Tabla 4.3. Reparto de créditos en la titulación.

A continuación se muestran las asignaturas impartidas en cada uno de los cursos con su correspondiente número de créditos.

Primer curso

El primer curso consta de 60 créditos troncales, repartidos en 2 cuatrimestres, 30 en el primero y 30 en el segundo. Las 10 asignaturas impartidas en el primer curso son las siguientes:

Asignatura	Créditos	Cuatrimestre
Algebra Lineal	6	1
Cálculo I	6	1
Mecánica Fundamental	6	1
Química I	6	1
Fundamentos Informática	6	1
Geometría	6	2
Cálculo II	6	2
Termodinámica Fundamental	6	2
Química II	5	2

Expresión Gráfica	7,5	2

Tabla 4.4. Asignaturas del primer curso

Segundo curso

En el segundo curso consta de 11 asignaturas troncales que suman un total de 57 créditos (31,5 en el primer cuatrimestre y 25,5 en el segundo). También hay un bloque optativo de 3 créditos en el segundo cuatrimestre.

Asignatura	Créditos	Cuatrimestre
Electromagnetismo	6	1
Métodos Numéricos	4,5	1
Materiales	4,5	1
Ecuaciones Diferenciales	6	1
Informática	4,5	1
Mecánica	6	1
Economía Empresa	6	2
Estadística	6	2
Dinámica Sistemas	4,5	2
Proyecto I	3	2
Teoría Maquinas Mecanismos	6	2
Optativas	3	2

Tabla 4.5. Asignaturas del segundo curso

Tercer curso

En el tercer curso se imparten 12 asignaturas troncales que suman un total de 60 créditos. Tanto el primer cuatrimestre como el segundo constan de 30 créditos repartidos en 6 asignaturas.

Asignatura	Créditos	Cuatrimestre
Tecnología Medio Ambiente y Sostenibilidad	6	1
Termodinámica	6	1
Electrotecnia	6	1
Mecánica Medios Continuos	4,5	1
Técnicas Estadísticas Calidad	3	1
Tecnología Selección Materiales	4,5	1
Mecánica Fluidos	6	2
Organización Gestión	4,5	2
Resistencia Materiales	6	2
Proyecto II	3	2
Maquinas Eléctricas	6	2
Optimización Simulación	4,5	2

Tabla 4.6. Asignaturas del tercer curso

Cuarto curso

El cuarto año está formado por un total de 60 créditos, de los cuales 30 son troncales, 18 son optativos, y 12 corresponden al Trabajo de Final de Grado (TFG). El primer cuatrimestre consta de 5 asignaturas que suman 30 créditos, y los 30 créditos restantes se imparten en el segundo cuatrimestre con el bloque optativo y el TFG.

Asignatura	Créditos	Cuatrimestre
Gestión Proyectos	6	1

Electrónica	7,5	1
Sistemas Fabricación	4,5	1
Termotecnia	6	1
Control Automático	6	1
TFG	12	2
Optativas	18	2

Tabla 4.7. Asignaturas del cuarto curso

5. Técnicas estadísticas

En este capítulo se mostrarán brevemente las diferentes técnicas estadísticas que se han aplicado sobre las variables. La función de dichas pruebas es dar una descripción numérica de los datos analizados y determinar qué variables discriminan correctamente entre los grupos creados para poder generar un modelo que clasifique de la manera más precisa posible.

5.1. Estadística descriptiva

A continuación se presentan los estadísticos empleados a lo largo de este proyecto.

Media aritmética

Es el resultado de sumar el valor de todos los datos y dividir el resultado entre el número total de datos. Su fórmula es la siguiente:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

Varianza

Es la media aritmética del cuadrado de las desviaciones respecto a la media aritmética de una muestra. Se calcula de la siguiente manera:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{N}$$

Asimetría

La asimetría es una medida que permite determinar si los datos se distribuyen de manera uniforme alrededor de la media aritmética, o por el contrario se concentran por encima o por debajo de la media. Se calcula mediante el coeficiente de asimetría de Fisher:

$$g_1 = \frac{\frac{1}{n}\sum\left(X_i - \overline{X}\right)^3 \cdot n_i}{\left(\frac{1}{n}\sum\left(X_i - \overline{X}\right)^2 \cdot n_i\right)^{\frac{3}{2}}}$$

Donde X_i representa cada uno de los valores, $\overline{X}$ la media de la muestra, n_i la frecuencia de cada valor, y n el número total de datos.

Si g_1 = 0 se puede afirmar que la curva es simétrica, por lo que existe aproximadamente la misma cantidad de datos en los dos lados de la media. A la práctica, se consideran distribuciones simétricas cuando el valor de $g_1 = \pm 0{,}5$.

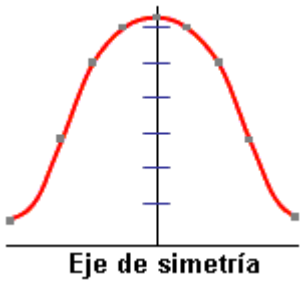

Figura 5.1. Distribución simétrica

Si g_1 > 0 indica que los valores tienden a concentrarse más a la izquierda de la media. Se dice que la curva es asimétricamente positiva.

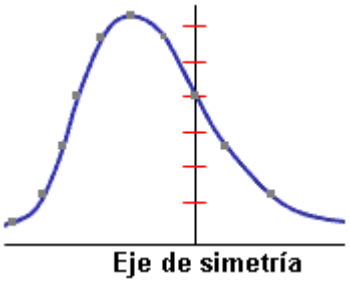

Figura 5.2. Distribución asimétricamente positiva

Si g_1 < 0 la curva es asimétricamente negativa y los valores tienden a concentrarse más en la parte de la derecha de la media.

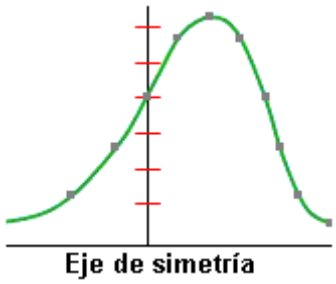

Figura 5.3. Distribución asimétricamente negativa

Curtosis

El coeficiente de curtosis indica el grado de concentración que presentan los datos en la región central de la distribución. Se calcula de la siguiente manera:

$$g_2 = \frac{\frac{1}{n}\sum\left(X_i - \overline{X}\right)^4 \cdot n_i}{\left(\frac{1}{n}\sum\left(X_i - \overline{X}\right)^2 \cdot n_i\right)^2} - 3$$

Donde X_i representa cada uno de los valores, $\overline{X}$ la media de la muestra, n_i la frecuencia de cada valor, y n el número total de datos.

Si g_2 = 0 los datos presentan una concentración normal alrededor de la región central de la distribución y se dice que la distribución es mesocúrtica. La tolerancia para aceptar que la distribución es mesocúrtica es $g_2 = \pm 0{,}5$

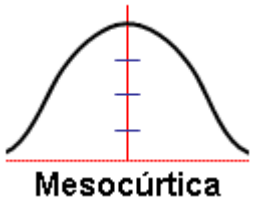

Figura 5.4. Distribución mesocúrtica

Si g_2 > 0 la distribución es leptocúrtica, y significa que existe una gran concentración de valores alrededor de la zona central de la distribución.

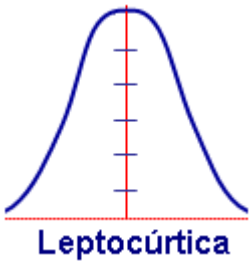

Figura 5.5. Distribución leptocúrtica

Si g_2 < 0 existe una baja concentración alrededor de la zona central de la distribución y se dice que la distribución es platicúrtica.

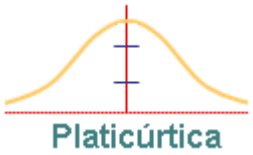

Figura 5.6. Distribución platicúrtica

5.2. Otras técnicas

Las técnicas estadísticas consideradas para su uso en este proyecto se pueden dividir en dos grandes grupos: las paramétricas y las no paramétricas. Las pruebas paramétricas tienen una mayor capacidad para detectar una relación en caso de que ésta exista. Por contra, exige que se cumplan una serie de requisitos para su aplicación. Las condiciones que se deben cumplir son las siguientes:

1. Que las variables dependientes estén medidas en una escala numérica.
2. Que los datos de las variables dependientes sigan una distribución normal.
3. Que las variables dependientes tengan homocedasticidad, es decir, igualdad de varianzas entre los grupos.

En caso de que alguno de estos requisitos no se cumpla, es necesario recurrir a las pruebas no paramétricas, las cuales tienen una capacidad menor a la hora de detectar relaciones entre los datos, pero los requisitos exigidos para su aplicación son menores, hecho que permite analizar un abanico de variables más amplio.

Existen distintas pruebas para comprobar si las variables cumplen estos supuestos de normalidad y de homocedasticidad. En este proyecto se ha empleado el test de Kolmogorov - Smirnov para determinar si los datos siguen una distribución normal, y el test de Levene para saber si existe igualdad de varianzas entre los grupos.

5.2.1. Prueba de Kolmogorov - Smirnov

Ésta constituye una prueba de bondad de ajuste, y se emplea para comprobar si los datos se ajustan a alguna distribución teórica específica, que en el caso de este proyecto es una distribución normal. Dicha prueba es aplicable a cualquier tamaño muestral, y el único requisito que exige para su aplicación es que los datos de la muestra sean continuos. El primer paso antes de realizar la prueba es fijar una hipótesis nula (H_0) y una hipótesis alternativa (H_a). Dichas hipótesis son las siguientes:

- H_0: Los datos analizados siguen una distribución normal.
- H_a: Los datos analizados no siguen una distribución normal.

El fundamento de la prueba de Kolmogorov-Smirnov se basa en comparar la frecuencia acumulada teórica con la frecuencia acumulada observada y determinar el punto de mayor divergencia.

El estadístico que se emplea es el siguiente:

$$D = \max\left|F_n(x_i) - F_0(x_i)\right|$$

Donde $F_n(x_i)$ es un estimador de la probabilidad de obtener valores menores o iguales que x_i, y $F_0(x_i)$ es la probabilidad de obtener valores menores o iguales que x_i cuando la hipótesis nula (H_0) es cierta (cuando siguen una distribución normal en el caso de este proyecto). Para la aplicación del estadístico D, es necesario haber ordenado de menor a mayor los datos de la muestra.

El criterio que se sigue para rechazar o no la hipótesis nula puede ser de dos tipos. El primero de ellos es el siguiente:

- Si $D \leq D_\alpha \Rightarrow$ Se acepta la hipótesis nula.
- Si $D > D_\alpha \Rightarrow$ Se rechaza la hipótesis nula.

Donde D_α es un valor tabulado para un cierto valor de significación α, que se define como la probabilidad de rechazar la hipótesis nula siendo ésta cierta.

El segundo criterio que se emplea para la toma de decisiones acerca de las hipótesis es el que se lleva a cabo mediante el p-valor asociado al estadístico D. La mayoría de programas estadísticos realizan automáticamente el cálculo del p-valor. Cuanto mayor es éste, indica que siendo la hipótesis nula H_0 cierta, era esperable obtener unos resultados como los observados. Por el contrario, un p-valor pequeño indica que es muy difícil obtener unos resultados como los observados siendo la hipótesis nula cierta.

La regla de decisión que se emplea en este caso es:

- Si $p - valor \geq \alpha \Rightarrow$ Se acepta la hipótesis nula.
- Si $p - valor < \alpha \Rightarrow$ Se rechaza la hipótesis nula.

Donde α es el nivel de significación, que en este proyecto será del 5 %.

5.2.2. Prueba de independencia de Chi - Cuadrado

La distribución de Chi - Cuadrado es empleada para distintas técnicas de análisis estadístico, como pueden ser pruebas de bondad de ajuste para saber si una muestra se adapta a una determinada distribución teórica, o pruebas de homogeneidad para saber si varias muestras cualitativas provienen de una misma población.

En el caso que ocupa el presente proyecto se ha empleado la distribución de Chi - Cuadrado para realizar una prueba de independencia. El objetivo de dicha prueba es determinar si dos variables cualitativas están relacionadas entre sí.

De igual modo que en el test de Kolmogorov - Smirnov, se debe fijar en primer lugar una hipótesis nula (H_0) y una hipótesis alternativa (H_a). Para esta prueba, las dos hipótesis son las siguientes:

- H_0: Las dos variables analizadas son independientes.
- H_a: Las dos variables analizadas son dependientes.

Una vez fijadas las hipótesis, se debe construir una tabla de contingencia de la siguiente manera:

	x_1	x_2	...	x_m	Total
y_1	n_{11}	n_{12}		n_{1m}	$\sum_{m=1}^{m} n_{1m}$
y_2	n_{21}	n_{22}		n_{2m}	$\sum_{m=1}^{m} n_{2m}$
...					
y_k	n_{k1}	n_{k2}		n_{km}	$\sum_{m=1}^{m} n_{km}$
Total	$\sum_{k=1}^{k} n_{k1}$	$\sum_{k=1}^{k} n_{k2}$		$\sum_{k=1}^{k} n_{km}$	$\sum_{m=1}^{m}\sum_{k=1}^{k} n_{km}$

Tabla 5.1. Tabla de contingencia.

Donde x_i son los distintos valores que puede adoptar la variable X, e y_i son los distintos valores que puede adoptar la variable Y. El estadístico de la prueba de Chi Cuadrado es el siguiente:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

donde n_{ij} frecuencias observadas en los datos a analizar, y e_{ij} son las frecuencias teóricas esperadas, que se calculan como se indica a continuación:

$$e_{ij} = \frac{\sum_{j=1}^{m} n_{ij} \cdot \sum_{i=1}^{k} n_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}}$$

El valor de χ^2 calculado se debe de comparar con el percentil de la distribución de Chi Cuadrado con (m-1)·(k-1) grados de libertad y para un determinado nivel de significación,

que en el caso de este proyecto será del 5 %. Si el valor calculado es inferior al valor crítico obtenido con la distribución, se acepta la hipótesis nula y se asume que las variables X e Y analizadas son independientes. Por el contrario, si el valor es superior, se rechaza la hipótesis nula y se asume que existe una relación entre las variables.

Del mismo modo que en la prueba de Kolmogorov - Smirnov, el criterio para rechazar o no la hipótesis nula puede estar basado en el p - valor, que como ya se ha dicho en otros capítulos, es un valor calculado y proporcionado por la mayoría de programas estadísticos. Para un nivel de significación α ,la regla es la siguiente:

- Si $p-valor \geq \alpha \Rightarrow$ Se acepta la hipótesis nula y se asume que las variables son independientes.
- Si $p-valor < \alpha \Rightarrow$ Se rechaza la hipótesis nula y se asume que existe una relación entre las variables.

5.2.3. Test U de Mann - Whitney

Puede considerarse la alternativa no paramétrica a la prueba de la t - student y permite comprobar si dos muestras independientes proceden o no de una misma población. El único requisito para su aplicación es que los datos estén medidos en algún tipo de escala ordinal o que sean continuos.

Del mismo modo que los test explicados en este capítulo, en primer lugar se debe fijar una hipótesis nula (H_0) y una hipótesis alternativa (H_a). En este caso son las siguientes:

- H_0: Las dos muestras, de tamaños n_1 y n_2 respectivamente, proceden de la misma población.
- H_a: Los valores de los datos de una de las muestras difieren de los de la otra.

Para realizar el contraste, en primer lugar se deben ordenar conjuntamente las observaciones de las dos muestras, de menor a mayor, y se les asigna un rango que va de 1 a $n_1 + n_2$. En el caso de valores iguales (ligas o empates), se les asigna un rango promedio. Si las dos muestras proceden de una misma población, los rangos se distribuyen aleatoriamente y la tendencia central de los mismos es similar. Por el contrario, si las dos muestras provienen de poblaciones distintas, los rangos se distribuyen con tendencias diferentes y los valores R_1 y R_2 difieren.

Con los rangos R_1 y R_2 calculados, se procede a calcular el estadístico U, que se define como el mínimo entre U_1 y U_2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Donde n_1 y n_2 son los tamaños de las muestras 1 y 2 respectivamente, y R_1 y R_2 es la suma de rangos de las observaciones de las muestras 1 y 2 respectivamente.

La distribución del estadístico U para muestras grandes (entiéndase por muestras grandes más de 20 observaciones) se aproxima bastante bien a una distribución normal de parámetros:

$$\mu_u = \frac{n_1 n_2}{2}$$

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Donde μ_u y σ_u son la media y la desviación estándar de U si la hipótesis nula es cierta. Con estos dos parámetros calculados, se obtiene el estadístico Z:

$$Z = \frac{U - \mu_u}{\sigma_u}$$

La zona de rechazo de la hipótesis nula se encuentra en las dos colas si la hipótesis alternativa es bilateral o en una si es unilateral. Si la probabilidad de obtener ese valor para el estadístico Z es menor que el nivel de significación α, se rechaza la hipótesis nula.

5.2.4. Test de Kruskal Wallis

El test de Kruskal Wallis es una prueba no paramétrica que se emplea para saber si r muestras independientes proceden de una misma población o no. Se puede considerar una extensión de la prueba U de Mann - Whitney para tres o más grupos, y al igual que en esta última, se debe fijar en primer lugar una hipótesis nula (H_0) y una hipótesis alternativa (H_a):

- H_0 : Las r muestras proceden de la misma población.
- H_a : Las observaciones de por lo menos una de las muestras procede de una población distinta.

Para realizar el contraste se deben ordenar los datos de las muestras de manera conjunta de menor a mayor, y asignarles un rango, que va desde 1 hasta N, que es el número total de observaciones. Para cada una de las r muestras se debe realizar el sumatorio de los rangos (R_m). A continuación se aplica el estadístico de contraste, que es el siguiente:

$$H = \frac{12}{N(N+1)} \sum_{m=1}^{r} \frac{R_m^2}{n_m} - 3(N+1)$$

Donde N número total de datos de todas las muestras, R_m es el sumatorio de los rangos de cada una de las r muestras, y n_m es el número de datos de cada una de las r muestras.

Cuando varias observaciones de la misma o de distintas muestras son iguales y se les asigna el mismo rango es necesario aplicar un factor de corrección a H. El efecto de este factor de corrección es elevar ligeramente el valor de H. Dicho factor es el siguiente:

$$H' = \frac{H}{1 - \frac{\sum_{i=1}^{g} \left(t_i^3 - t_i\right)}{N^3 - N}}$$

En esta expresión, g es el número de rangos que se repiten y t_i es el número de veces que se repite el rango i.

Si los tamaños muestrales son grandes (recordemos que entendemos por grande más de 20 observaciones), el estadístico H se aproxima a una distribución Chi - Cuadrado con r-1 grados de libertad. Por lo tanto, para un nivel de significación α, la regla de decisión es la siguiente:

- Si $H > \chi^2_{r-1,1-\alpha} \Rightarrow$ Se rechaza la hipótesis nula H_0.
- Si $H \leq \chi^2_{r-1,1-\alpha} \Rightarrow$ Se acepta la hipótesis nula H_0.

5.2.5. Regresión lineal

La regresión lineal es una técnica estadística empleada para estudiar la relación entre dos variables cuantitativas. Las variables dependientes son aquellas que se desean predecir, mientras que las variables independientes (o explicativas) son las empleadas para predecir el valor de la dependiente mediante una ecuación lineal. Se puede hablar de regresión simple y de regresión múltiple. La regresión simple es aquella que está formada por únicamente dos variables, una dependiente y una independiente, mientras que la regresión múltiple es aquella formada por una variable dependiente y dos o más variables independientes. La expresión general de la regresión lineal múltiple es la siguiente:

$$\hat{Y} = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k + U$$

Siendo Y el valor estimado de la variable dependiente, β los coeficientes que se deberán hallar, y U es el término de perturbación o error. Observando la expresión, se puede deducir que los coeficientes β marcarán el aumento de Y por el aumento unitario de la correspondiente variable independiente.

El objetivo consiste en asignar valores a los coeficientes β, de manera que la distancia entre lo estimado y lo observado sea la mínima posible. Con el fin de poder determinar las propiedades de los estimadores y de emplear un método óptimo para encontrar su valor, se deben cumplir las siguientes hipótesis:

1. Linealidad: los valores de la variable dependiente vienen dados por la siguiente expresión:

   - $Y = \beta \cdot X + U$

2. Homocedasticidad: igualdad de varianzas en los términos de error U.

3. Independencia: los términos de error no están correlacionados entre sí.

4. Normalidad: las perturbaciones o términos de error siguen una distribución normal cuyo valor esperado es 0.

5. No hay errores de medida en las variables explicativas X.

La estimación del modelo, que consiste en asignar un valor numérico a los parámetros β es calculada por el programa SPSS. Existen distintos métodos para realizarla, pero dos de los

más conocidos son el método de los mínimos cuadrados ordinarios (MCO) y el método de máxima verosimilitud (MV).

El método de mínimos cuadrados ordinarios trata de minimizar la suma de los cuadrados de los residuos con respecto a los parámetros β, donde el vector de residuos *u* viene dado por la expresión $u = Y - \hat{Y} = Y - X * \hat{B}$. Resolviendo el problema de optimización se obtiene la siguiente expresión para el vector de coeficientes β:

$$\hat{\beta} = \left(\hat{\beta}_0 \quad \hat{\beta}_1 \ldots \quad \hat{\beta}_k\right)^T = \left(X^T X\right)^{-1} X^T Y$$

Donde X es la matriz de variables explicativas e Y es la matriz de valores observados.

Si bien el método de los mínimos cuadrados ordinarios (MCO) trataba de minimizar la suma de los cuadrados de los residuos, el método de máxima verosimilitud trata de maximizar la probabilidad de obtener una muestra como la que se dispone. Para ello, se debe maximizar la función de densidad conjunta del vector aleatorio U, que tal y como se ha fijado en las hipótesis, sigue una distribución normal. Debido a que U se puede expresar en función de Y, X, β, resolviendo dicho problema de maximización se obtiene que los valores del vector de coeficientes β se calculan de la siguiente manera:

$$\hat{\beta} = \left(\hat{\beta}_0 \quad \hat{\beta}_1 \ldots \quad \hat{\beta}_k\right)^T = \left(X^T X\right)^{-1} X^T Y$$

Se observa que los estimadores obtenidos por ambos métodos son los mismos, ya que si las hipótesis marcadas son ciertas, ambos dan una solución óptima.

El programa SPSS permite seleccionar el método por el cual se van introduciendo las variables independientes en el modelo. Los más relevantes son los siguientes:

- **Método Introducir:** todas las variables preseleccionadas participan en la regresión y se introducen todas de un solo paso.
- **Método hacia atrás:** se introducen todas las variables preseleccionadas en el modelo y se van eliminando en pasos sucesivos. El orden de consideración para la exclusión de variables es de menos a más correlación parcial con la variable dependiente. Si la variable cumple los criterios de eliminación es eliminada.
- **Método hacia adelante:** se van introduciendo las variables preseleccionadas de una en una y por pasos. El orden de consideración de las variables de entrada

es establecido por su correlación parcial con la variable dependiente, teniendo prioridad para entrar aquellas con una correlación mayor. Si la variable cumple con los criterios de entrada, será introducida en el modelo.

Tras contrastar los resultados obtenidos con los distintos métodos se concluye que el método hacia adelante es el que ofrece unos resultados más precisos, por lo que será el empleado en el presente proyecto. Los criterios de entrada de las variables en dicho método están basados en el estadístico F de Snedecor, con el cual se contrasta la significación conjunta del modelo. Si el p-valor calculado por el programa es menor o igual a 0,05, la nueva variable se introduce en el modelo.

Bondad de ajuste de la regresión lineal: Coeficiente de R Cuadrado

Con el modelo obtenido en cada uno de los pasos, resulta interesante medir la bondad de ajuste del mismo. Para ello, el programa SPSS utiliza el coeficiente de determinación de R Cuadrado. Para explicar el cálculo de R Cuadrado, es necesario introducir algunos conceptos como el SCT. El SCT (Suma de Cuadrados Totales) representa la suma de cuadrados de las desviaciones de los datos observados respecto a su media aritmética. Su fórmula es la siguiente:

$$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

El SCT representa la variabilidad total de la variable dependiente, y está compuesto por dos partes, la suma de cuadrados de la regresión (SCR) y la suma de cuadrados de los errores (SCE). Por lo tanto $SCT = SCR + SCE$, o lo que es lo mismo pero expresado de otra manera:

$$SCT = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^{\,2}$$

Donde el primer término de la derecha de la igualdad representa el SCR, y el segundo el SCE. Por lo tanto, de la variabilidad total de la variable dependiente, hay un parte que es explicada por el modelo, que es el SCR, y otra que no, que es la suma del cuadrado de los errores (SCE). El coeficiente de bondad de ajuste de la R Cuadrado es el cociente entre la parte de variabilidad explicada por el modelo y la variabilidad total:

$$R\ \ Cuadrado = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Dicho de otro modo, el parámetro de R Cuadrado indica la proporción de variabilidad e la variable dependiente explicada por a regresión lineal.

5.2.6. Regresión logística binaria

La regresión logística binaria es un tipo de regresión que su utiliza para predecir el resultado de una variable dicotómica (que puede adoptar dos valores distintos) en función de distintas variables independientes o predictoras. Esta técnica permite modelar la probabilidad de que suceda un determinado evento en función de los valores de distintas variables, que pueden ser cuantitativas o cualitativas. Dicha probabilidad debe estar entre 0 y 1, y a distintos valores de las variables predictoras ésta ha de ser diferente.

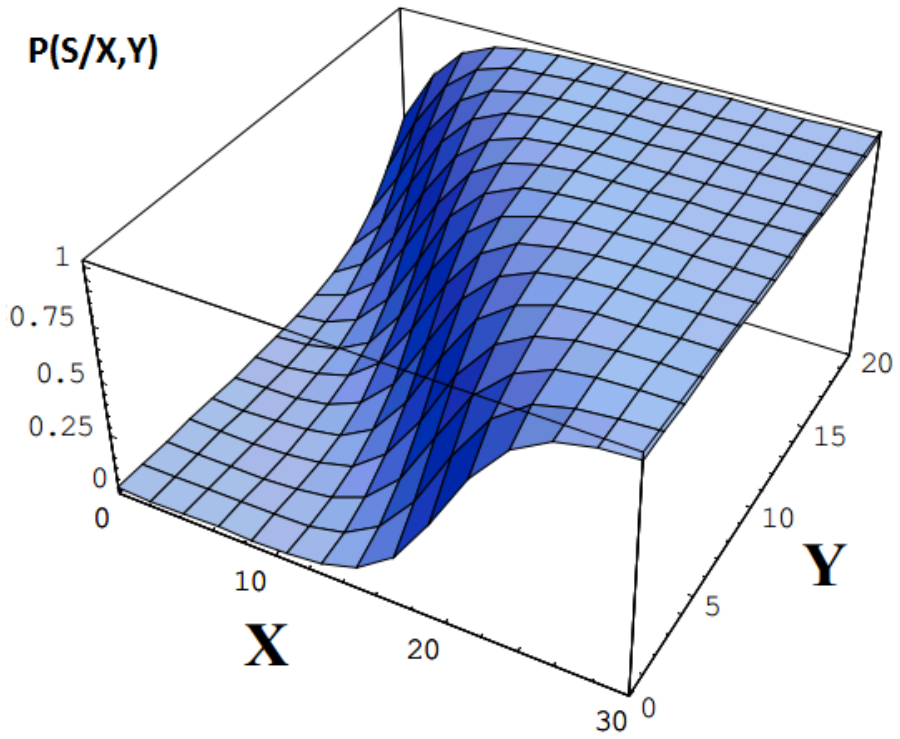

Figura 5.7. Probabilidad de que S = 1 en función de X e Y

Dada una variable S que puede adoptar valores 0 y 1, y un individuo con un conjunto k de variables independientes $(X_1, X_2, X_3 ... X_K)$, la probabilidad de que S=1 dada por el modelo logístico es la siguiente:

$$P(S=1/X_1,X_2,...X_K)=\frac{1}{1+e^{(-\beta_0-\beta_1 x-\beta_2 x_2-...-\beta_k x_k)}}$$

Donde los coeficientes $\beta\,(\beta_0,\beta_1,\beta_2...\beta_K)$ se calculan usando una estimación de máxima verosimilitud, que es aquella que otorga a los parámetros el valor que haga máxima la probabilidad de obtener una muestra como la observada. Con ello se consigue una función de enlace tal que $P\left[S=1/X_1,X_2...X_k\right]=p(X_1,X_2...X_k;\beta)$.

Para hallar la expresión del modelo de regresión logístico binario, es necesario definir el concepto de ODDS (ratio del riesgo), que no es más que el cociente entre que ocurra un determinado evento y que no ocurra. Viene dado por la siguiente expresión:

$$ODDS=\frac{p}{1-p}$$

Introduciendo en la fórmula de ODDS la expresión de probabilidad se obtiene la siguiente expresión:

$$ODDS=\frac{P(S=1/X_1,X_2,...X_K)}{1-P(S=1/X_1,X_2,...X_K)}=\frac{p(X_1,X_2...X_k;\beta)}{1-p(X_1,X_2...X_k;\beta)}=e^{(\beta_0+\beta_1 x-\beta_2 x_2+...+\beta_k x_k)}$$

Aplicando logaritmos neperianos en la expresión anterior, se obtiene una expresión lineal para el modelo:

$$Logit\left[P(S=1)\right]=Ln\left[\frac{P(S=1/X_1,X_2,...X_K)}{1-P(S=1/X_1,X_2,...X_K)}\right]=\beta_0+\beta_1 x+\beta_2 x_2+...+\beta_k x_k$$

Al ser una ecuación lineal, cualquier variación de los parámetros dará un resultado distinto de la función Logit.

Resulta interesante introducir el concepto de Odds-Ratio, que se define como el cociente entre los Odds asociados al realizar un incremento unitario de la variable X_i, y los Odds que tenía antes de realizar el incremento:

$$ODDS-Ratio=\frac{ODDS\,2}{ODDS\,1}=\exp(\beta_i)$$

De dicha expresión se deduce que si el coeficiente β_i es cercano a cero (dando lugar a un ODDS-Ratio cercano a uno), una variación de la variable X_i asociada al coeficiente no tendrá efectos sobre la variable dependiente S.

6. Resultados

6.1. Estadística descriptiva

Dentro de este apartado se encuentran todos aquellos datos de carácter general de cada una de las titulaciones, así como la estadística descriptiva de las distintas asignaturas de la titulación Grado en Ingeniería en Tecnologías Industriales impartida en la ETSEIB. El objetivo de dichas estadísticas es realizar una primera aproximación numérica a los datos disponibles y detectar los casos más extremos o más anómalos para poder ser analizados posteriormente. La aplicación de la estadística descriptiva a las distintas asignaturas se segmentará por cursos, y por parámetros calculados, mostrando en primer lugar información como la nota media de cada asignatura, la nota mínima y máxima, y el número de aprobados y suspendidos; y en segundo lugar, estadísticos como la varianza, la asimetría, y la curtosis de cada asignatura con el fin de intuir el nivel de centralización de los datos y la forma que tiene la distribución. También se incluirá una sección para mostrar la evolución temporal de la nota media de las asignaturas a lo largo de los distintos años de los que se poseen datos. De este modo se detectará si existe algún cambio brusco o alguna tendencia en alguna de las asignaturas.

6.1.1. Estadísticas generales de los grados impartidos en la ETSEIB

Tal y como ya se ha dicho anteriormente, en la ETSEIB se cursan tres titulaciones de Grado. En los datos que se dispone, el reparto de alumnas en las distintas titulaciones es el siguiente:

Titulación	Nº Alumnos	% del Total
Grado en Ingeniería en Tecnologías Industriales	1820	79,30%
Grado en Ingeniería Química	304	13,25%
Grado en Ingeniería de Materiales	171	7,45%
Total	**2295**	

Tabla 6.1. Reparto de los alumnos en los distintos Grados.

Se puede apreciar como hay una titulación con un número mayoritario de alumnos, que es el Grado en Ingeniería en Tecnologías Industriales. Prácticamente un 80 % de los alumnos registrados en la base de datos cursan esta titulación, motivo por el cual el presente proyecto se centrará en dicha titulación. En segundo lugar, pero con un porcentaje mucho menor, se

encuentra Grado en Ingeniería Química, que cuenta con 304 estudiantes (un 13,25 % del total de estudiantes). Por último, la titulación menos cursada es Grado en Ingeniería de Materiales, con un total de 171 estudiantes que representa un 7,45 % del total.

Segmentando el número de estudiantes por sexo y por titulación, se distribuyen del siguiente modo:

Titulación	Hombres	Mujeres
Grado en Ingeniería en Tecnologías Industriales	1413 (77,64 %)	407 (22,36 %)
Grado en Ingeniería Química	189 (62,17 %)	115 (37,83 %)
Grado en Ingeniería de Materiales	133 (77,78 %)	38 (22,22 %)
Total	1735 (75,60 %)	560 (24,40 %)

Tabla 6.2. Número de hombres y mujeres por titulación

Se observa que un 75,60 % de los estudiantes son hombres y sólo el 24,40 % son mujeres. La titulación con más presencia de sexo femenino es la 753 (Grado en Ingeniería Química), en la cual el 37,83 % de los estudiantes son mujeres.

Como ya se ha comentado en capítulos anteriores, debido al número de datos este proyecto se centrará únicamente en el Grado en Ingeniería en Tecnologías Industriales. En el siguiente apartado se muestran algunas estadísticas generales de dicha titulación.

6.1.2. Estadísticas generales Grado en Ingeniería en Tecnologías Industriales

Se disponen de datos de 1820 alumnos cursando dicha titulación, cuyos años de ingreso en la ETSEIB se encuentran entre 2010 y 2013. En primer lugar, se realiza el cálculo de las notas promedio obtenidas en cada curso. Dichas notas medias se han calculado únicamente con los alumnos que han completado el curso. Por lo tanto, si un alumno repite una asignatura y al cuatrimestre siguiente la aprueba, la nota media será calculada con la nota de la asignatura aprobada.

	Nota media	N
Primero	6,35	1439
Segundo	6,56	609
Tercero	6,48	372
Cuarto	7,54	170

Tabla 6.3. Notas medias por cursos.

Se puede observar como a medida que se va pasando de curso la muestra con la que se ha calculado la media disminuye. Esto sucede debido a que, si se poseen datos hasta el primer cuatrimestre de 2014, únicamente la primera promoción de grado (alumnos que ingresaron en el año 2010) es la que debería haber completado cuarto yendo a curso por año. Se puede apreciar también como la nota promedio en cuarto es notablemente más alta que la del resto de cursos. Como se verá más adelante, la respuesta a este suceso está en el TFG, cuya nota media es un 8,73 y su peso es de 12 créditos.

Desglosando el promedio por sexo, se puede apreciar como apenas existe diferencia entre las medias de los hombres y de las mujeres.

	Hombres	Nº Hombres	Mujeres	Nº Mujeres
Promedio Primero	6,37	1097	6,30	343
Promedio Segundo	6,57	476	6,51	133
Promedio Tercero	6,49	272	6,45	100
Promedio Cuarto	7,50	127	7,64	43

Tabla 6.4. Promedios por cursos desglosados por sexo

En cuanto al número de asignaturas repetidas, a continuación se puede observar el promedio de cada uno de los cursos para los estudiantes que los han completado:

	Promedio Asignaturas Repetidas	Nº Estudiantes
Primero	2,49	1439
Segundo	1,01	609
Tercero	0,97	372
Cuarto	0,12	170

Tabla 6.5. Promedio de asignaturas repetidas por curso

Se puede apreciar una tendencia claramente descendente en el promedio de asignaturas suspendidas por curso. Es lógico pensar que a medida que van avanzando los cursos, a parte de la fase selectiva, se crea un filtro natural que deriva en el abandono de los estudios por parte de aquellos alumnos con peores resultados. No obstante, se sospecha que los promedios de los últimos cursos pueden estar ligeramente alterados debido al corto periodo

de tiempo del que se disponen datos, ya que a medida que avanzan los cursos, sólo van llegando los mejores. Explicado de otro modo, de las asignaturas de primer curso se poseen datos de estudiantes que han ingresado en 2010, 2011, 2012 y 2013; de las asignaturas de segundo curso los datos pertenecen a estudiantes ingresados en 2010, 2011, 2012 y los mejores de 2013, que son aquellos que han completado primero en un año. Para las asignaturas de tercero, los datos son aquellos de aquellos estudiantes que han ingresado en 2010, 2011, y los mejores de 2012. En definitiva, debido a los pocos años que lleva implantado el Plan Bolonia, los datos extraídos todavía no se encuentran en una fase estacional para los últimos cursos, hecho que puede producir una pequeña distorsión de los resultados. Para comprobar si esta hipótesis es cierta o no, se muestra a continuación el promedio de asignaturas repetidas por año de ingreso en la ETSEIB (sólo para los estudiantes que han completado los cursos):

Año de Ingreso	Promedio Repetidas Primero	Promedio Repetidas Segundo	Promedio Repetidas Tercero	Promedio Repetidas Cuarto
2010	2,84	1,03	1,22	0,12
2011	1,87	1,33	0,56	
2012	2,71	0,53		
2013	1,78			

Tabla 6.6. Promedio de asignaturas repetidas por año de ingreso del estudiante

Se observa que la hipótesis formulada no es cierta, ya que no se puede afirmar que exista una tendencia descendente en los promedios de las asignaturas repetidas en los distintos años. Sin embargo sí que se refuerza para todos los años de ingreso (salvo 2013 que apenas se tienen datos) que el promedio de asignaturas repetidas disminuye a medida que van avanzando los cursos. También es interesante remarcar que las variaciones en los promedios de asignaturas repetidas son bastante significativos. Puede interesar observar también qué sucede con los promedios de los distintos cursos por año de ingreso:

Año de Ingreso	Promedio Primero	Promedio Segundo	Promedio Tercero	Promedio Cuarto
2010	6,35	6,50	6,44	7,54
2011	6,43	6,52	6,54	7,13
2012	6,21	6,73		

2013	5,85			

Tabla 6.7. Notas promedio por cursos en función del año de ingreso del estudiante

Los promedios de los alumnos que han aprobado los distintos cursos se mantienen en una línea regular, ya que no hay variaciones especialmente grandes. Cabe destacar que no necesariamente los promedios más altos coinciden con aquellos años de ingreso en que el promedio de asignaturas repetidas es más alto, ya que dicho promedio es calculado cuando las notas de todas las asignaturas están aprobadas.

6.1.3. Análisis asignaturas de primer curso

Nota máxima, nota mínima y media

El primer curso de Grado en Ingeniería en Tecnologías Industriales cuenta con diez asignaturas repartidas en dos cuatrimestres. En la siguiente tabla se muestran las notas medias, máximas, mínimas y el número de aprobados y suspensos para cada una de las asignaturas.

Asignatura	N	Nota Mínima	Nota Máxima	Media	Aprobados	Suspensos
Algebra Lineal	2.345	0,0	10,0	5,24	1.600	745 (31,77 %)
Cálculo I	2.266	0,0	10,0	5,40	1.604	662 (29,21 %)
Cálculo II	2.065	0,0	9,7	5,32	1.533	532 (25,76 %)
Expresión Gráfica	2.029	0,0	10,0	5,51	1.580	449 (22,13 %)
Fundamentos Informática	2.472	0,0	10,0	5,28	1.638	834 (33,74 %)
Geometría	2.053	0,0	10,0	5,56	1.572	481 (23,43 %)
Mecánica Fundamental	2.324	0,0	9,9	5,26	1.608	716 (30,81 %)
Química I	2.184	0,0	9,8	5,71	1.677	507 (23,21 %)
Química II	2.000	0,0	10,0	5,84	1.587	413 (20,65 %)
Termodinámica Fundamental	2.243	0,0	9,7	5,05	1.547	696 (31,03 %)

Tabla 6.8. Estadísticas generales de las distintas asignaturas de primero.

Se observa en primer lugar que la nota mínima para todas las asignaturas es un 0,0 y que la máxima es un 10 salvo para cuatro de ellas: Cálculo II, Mecánica Fundamental, Química I y Termodinámica Fundamental, cuyas notas máximas son 9,7; 9,9; 9,8 y 9,7 respectivamente. Por lo tanto, desde 2010 que se poseen datos, no hay ningún estudiante que haya sacado un 10 en estas asignaturas. En cuanto a las notas medias, se puede observar que todas están por encima del 5, hecho que cambia si se calculan las notas medias incluyendo los

datos de Grado en Ingeniería Química y Grado en Ingeniería de Materiales. La asignatura de primero con una media más baja es Termodinámica Fundamental, con valor promedio de 5,05, y la asignatura de primero con una nota media más alta es Química II con un 5,84. Por otro lado, la asignatura con un porcentaje mayor de suspensos es Fundamentos de Informática, con un 33,74 %, y asignatura con un porcentaje menor es Química II con un 20,65 %.

Varianza, Asimetría y Curtosis

Mediante la observación de estos estadísticos se pretende localizar aquellas asignaturas que presentan una distribución un tanto extraña. En la siguiente tabla se muestra la varianza, la asimetría, y la curtosis para cada una de las asignaturas de primero:

Asignatura	Varianza	Asimetría	Curtosis
Algebra Lineal	3,61	-0,605	0,671
Cálculo I	3,19	-0,583	0,869
Cálculo II	3,09	-0,761	1,230
Expresión Grafica	4,30	-0,806	0,661
Fundamentos Informática	6,52	-0,482	-0,588
Geometría	2,64	-0,778	1,758
Mecánica Fundamental	2,65	-0,556	0,919
Química I	3,93	-0,869	1,103
Química II	3,19	-0,545	0,749
Termodinámica Fundamental	3,28	-0,403	0,148

Tabla 6.9. Varianza, asimetría y curtosis de las asignaturas de primero.

Se observa en primer lugar una varianza que destaca sobre las demás por tener un valor excesivamente alto, que es la de Fundamentos de Informática. Esto indica una alta dispersión de los datos, que se refleja también en el coeficiente de curtosis. Un valor negativo de este coeficiente es propio de las distribuciones platicúrticas, que son aquellas que presentan una baja concentración de datos en la región central de la distribución. Dado que la media de Fundamentos de Informática es de 5,28, se puede intuir que las notas de dicha asignatura son bastante extremas, es decir, o muy altas o muy bajas. Más adelante se realizará un análisis de dicha asignatura y de Expresión Gráfica, que también presenta una varianza con un valor más alto que el resto de asignaturas.

En cuanto a las asimetrías, se puede observar que son todas negativas. Este hecho significa que la cola de la distribución se alarga para valores más inferiores a la media.

Evolución temporal

Se desea observar cómo ha evolucionado la media de las asignaturas a lo largo del tiempo para ver las variaciones acontecidas. De las asignaturas de primero, se poseen datos desde 2010 hasta 2014. En la siguiente tabla se muestran las notas promedio de cada una de las asignaturas de primer curso:

Asignatura	2010	2011	2012	2013	2014
Algebra Lineal	5,24	5,43	5,55	4,76	5,58
Calculo I	5,54	5,67	5,30	5,15	4,87
Calculo II	5,66	6,02	4,73	5,07	4,68
Expresión Gráfica	5,42	5,78	5,29	5,59	5,23
Fundamentos Informática	4,63	5,38	5,42	5,68	6,10
Geometría	5,51	5,90	5,16	5,63	5,42
Mecánica Fundamental	5,38	4,99	5,04	5,68	4,75
Química I	5,96	5,15	5,65	6,16	5,18
Química II	6,03	6,02	6,00	5,38	5,76
Termodinámica Fundamental	5,07	5,16	5,20	4,74	5,16

Tabla 6.10. Evolución temporal de las notas de las asignaturas de primero.

También se ha graficado la evolución en dos gráficas, una para las asignaturas del primer cuatrimestre, y otra con las asignaturas del segundo, para permitir observar las variaciones de una manera más visual.

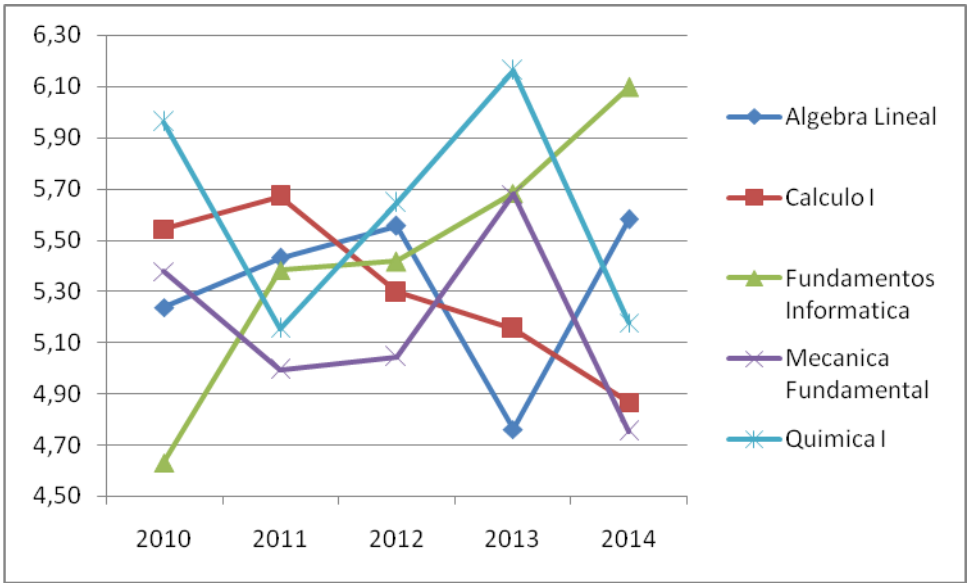

Figura 6.1. Evolución temporal de las asignaturas del primer cuatrimestre.

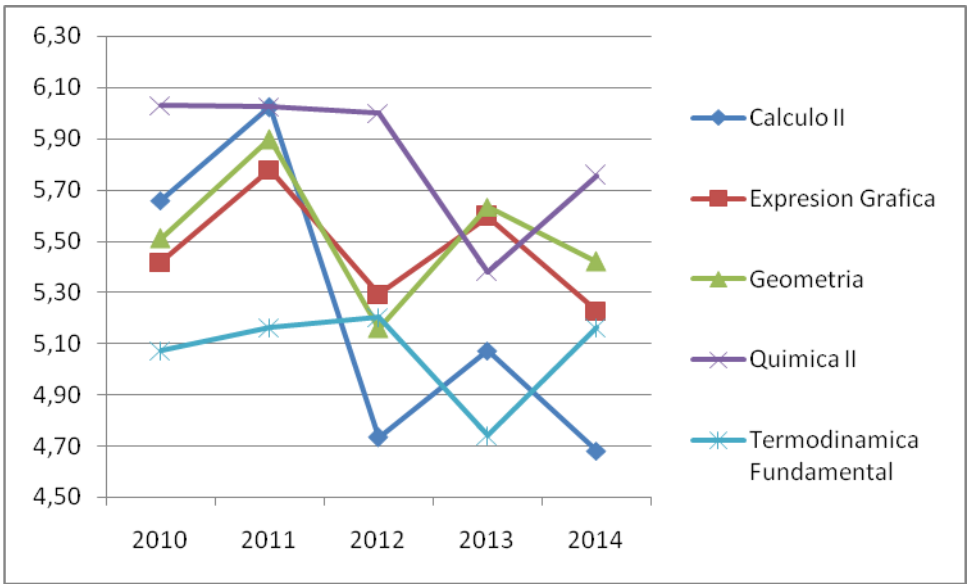

Figura 6.2. Evolución temporal de las asignaturas del segundo cuatrimestre.

La nota media más alta registrada corresponde a Química II con un valor de 6,16 en el año 2013, y la más baja corresponde a Fundamentos de Informática con un valor de 4,63 en el año 2010.

La variación positiva más grande que se ha experimentado es la de Algebra Lineal entre los años 2013 y 2014, con un incremento de 4,74 a 5,58 (un 17 %). Por el contrario, la bajada

de nota media más brusca es la encontrada en Cálculo II entre los años 2011 y 2012, con un descenso de un 21 %,bajando de 6,02 a 4,73.

En cuanto a las tendencias, se puede observar por un lado una tendencia ascendente en la nota de Fundamentos de Informática, que ha ido incrementando todos los años, pasando de 4,63 en 2010 hasta 6,10 en 2014. Por otro lado, se puede observar una tendencia negativa en la nota de Cálculo I desde 2011 hasta 2014, produciéndose un descenso de la nota desde 5,67 a 4,87.

Por último, resaltar que siempre que se ha producido un suspenso en la nota media de alguna asignatura, al año siguiente se ha producido un aumento en todos los casos, situando la nueva media en todos los casos por encima del 5. Este hecho puede ser una señal de reajustes realizados en la asignatura cuando el nivel de fracaso es elevado.

6.1.4. Análisis asignaturas de segundo curso

Nota máxima, nota mínima y media

El segundo curso de Grado en Ingeniería en Tecnologías Industriales cuenta con seis asignaturas en el primer cuatrimestre, y cinco en el segundo, sumando un total de 57 créditos sin contar el bloque de asignaturas optativas. Del mismo modo que se ha realizado con las asignaturas de primero, se muestran a continuación las notas medias, máximas, mínimas y el número de aprobados y suspensos para cada una de las asignaturas.

Asignatura	N	Nota Mínima	Nota Máxima	Media	Aprobados	Suspensos
Dinámica Sistemas	1.241	0,0	9,9	5,97	1.061	180 (14,5 %)
Economía Empresa	1.203	0,0	10,0	6,26	1.106	97 (8,06 %)
Ecuaciones Diferenciales	1.655	0,0	9,6	5,30	1.156	499 (30,15 %)
Electromagnetismo	1.700	0,0	9,5	5,13	1.143	557 (32,76 %)
Estadística	1.205	0,0	10,0	5,92	1.066	139 (11,54 %)
Informática	1.573	0,0	10,0	5,84	1.251	322 (20,47 %)
Materiales	1.611	0,0	9,7	5,25	1.100	511 (31,72 %)
Mecánica	2.058	0,0	10,0	4,33	1.055	1003 (48,74 %)
Métodos Numéricos	1.505	0,0	10,0	5,92	1.271	234 (15,55 %)
Proyecto I	1.087	0,0	10,0	8,37	1.085	2 (0,18 %)
Teoría Maquinas Mecanismos	1.304	0,0	10,0	5,62	966	338 (25,92 %)

Tabla 6.11. Estadísticas generales de las distintas asignaturas de segundo.

Tal y como sucede en primero, la nota mínima más baja es un cero para todas las asignaturas, y la más alta es un diez para todas salvo Dinámica de Sistemas, Ecuaciones Diferenciales, Electromagnetismo, y Materiales, cuyas notas máximas son 9,9; 9,6; 9,5; y 9,7 respectivamente. Todas las notas medias de las asignaturas están por encima del cinco menos una, que es Mecánica y cuyo valor es de 4,33. Del resto de medias, se puede destacar la de Proyecto I, que es 8,37 y es especialmente alta. En cuanto al porcentaje de suspensos, vuelve a destacar mecánica por tener un valor especialmente alto (un 48,74 %), lo que significa que prácticamente uno de cada dos alumnos repiten la asignatura. Por el contrario, destaca también Proyecto I, que desde 2011 hasta 2014 sólo han suspendido dos personas (un 0,18 %).

Varianza, Asimetría y Curtosis

Se procede de nuevo al estudio de estos estadísticos para detectar si existe alguna distribución de notas digna de ser analizada. Los valores calculados se encuentran en la siguiente tabla:

Asignatura	Varianza	Asimetría	Curtosis
Dinámica Sistemas	3,04	-1,111	2,429
Economía Empresa	1,55	-1,242	5,261
Ecuaciones Diferenciales	2,72	-0,747	1,883
Electromagnetismo	2,93	-0,674	0,963
Estadística	1,98	-0,966	4,286
Informática	3,84	-0,646	0,794
Materiales	2,03	-0,762	2,090
Mecánica	3,47	-0,306	0,552
Métodos Numéricos	2,93	-1,210	2,697
Proyecto I	1,07	-1,245	7,061
Teoría Maquinas Mecanismos	3,30	-0,236	0,225

Tabla 6.12. Varianza, asimetría y curtosis de las asignaturas de segundo

Tal y como sucedía en primero con Fundamentos de Informática, la asignatura de Informática de segundo vuelve a tener una varianza alta (3,84) comparado con el resto de asignaturas. Siguiendo a Informática, también tiene un valor de la varianza elevado Mecánica (3,47). Este hecho indica una alta variabilidad de los datos respecto a la media.

En cuanto a las asimetrías, se observan valores negativos en todas las asignaturas, por lo que la cola de la distribución queda por debajo de la media. No obstante, muchos valores son bastante cercanos a cero, por lo que no son distribuciones excesivamente asimétricas.

Observando los valores de la curtosis, hay uno que llama especialmente la atención, que es el de Proyecto I. Esto es propio de las funciones leptocúrticas, que son aquellas en que la distribución presenta una gran concentración de datos en su región central. En este caso, dado que el valor de la varianza es bajo, se puede afirmar que los datos están muy concentrados alrededor de la media. El mismo caso pero en menor escala se da en las asignaturas Economía de Empresa y Estadística, ya que se repite el mismo patrón de una curtosis alta y una varianza pequeña.

Evolución temporal

La evolución de las notas medias de las asignaturas de segundo entre los años 2011 y 2014 se muestran en la tabla y los gráficos que se muestran a continuación:

Asignatura	2011	2012	2013	2014
Dinámica Sistemas	6,56	5,85	5,99	5,62
Economía Empresa	6,53	6,31	6,11	6,12
Ecuaciones Diferenciales	5,76	5,26	5,21	4,87
Electromagnetismo	6,17	4,50	5,13	4,92
Estadística	5,92	6,16	5,88	5,42
Informática	6,40	5,43	5,80	5,89
Materiales	5,80	5,19	5,03	5,02
Mecánica	4,53	4,28	4,45	3,97
Métodos Numéricos	6,22	5,70	5,93	5,93
Proyecto I	8,44	8,26	8,52	8,27
Teoría Maquinas Mecanismos	6,37	5,58	5,70	4,89

Tabla 6.13. Evolución temporal de las notas de las asignaturas de segundo.

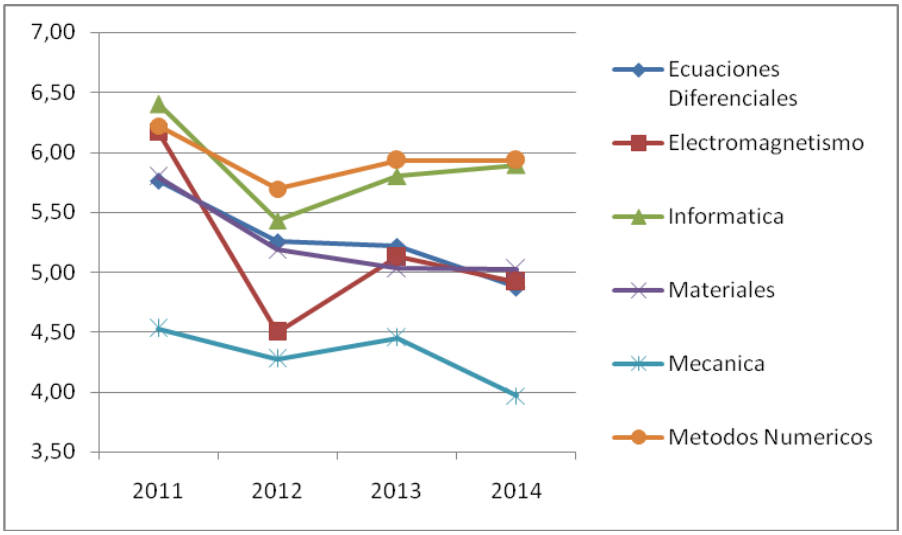

Figura 6.3. Evolución temporal de las asignaturas del primer cuatrimestre

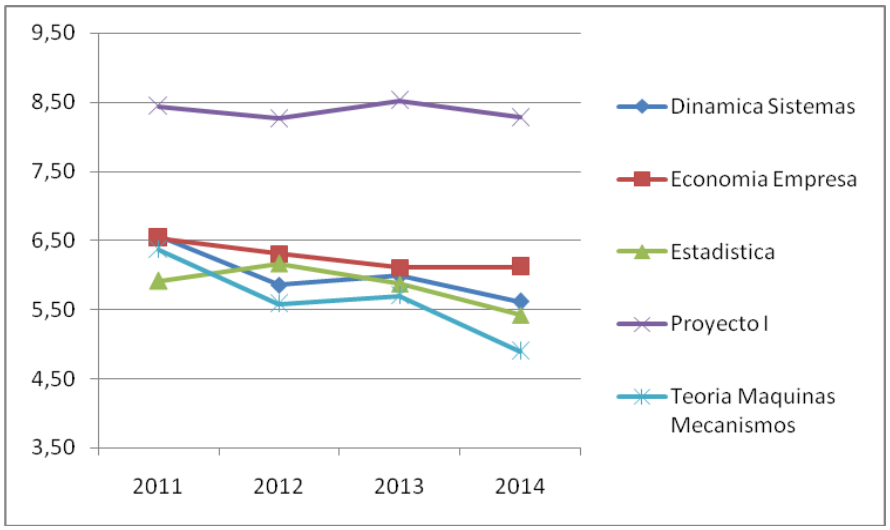

Figura 6.4. Evolución temporal de las asignaturas del segundo cuatrimestre

En primera instancia, se puede comprobar cómo la media de Proyecto I es notablemente superior al resto durante todos los años que se poseen datos (del 2011 al 2014). Asimismo, la media de Mecánica también destaca, pero en este caso por ser más significativamente más baja que el resto a lo largo de estos años. También pertenecen a estas dos asignaturas la media más alta registrada en estos años, y la más baja, siendo la más alta un 8,52 en Proyecto I en el 2013, y la más baja un 3,97 en Mecánica en el año 2014.

En cuanto a las variaciones más grandes producidas entre un año y otro se encuentran en la asignatura de Electromagnetismo. Entre 2011 y 2012 se encuentra el descenso más brusco de las notas medias de segundo curso, bajando de 6,17 a 4,50 (un 27,1 %). Sin embargo, entre 2012 y 2013 se observa la variación positiva más grande, subiendo la media de 4,50 hasta 5,13 (un 14,0 %).

En cuanto a las tendencias de las asignaturas, se puede observar como Ecuaciones Diferenciales y Materiales presentan una tendencia descendente entre los años 2011 y 2014. En el caso de Ecuaciones Diferenciales, la media ha descendido desde un 5,76 hasta un 4,87, y en el caso de Materiales ha bajado desde un 5,80 hasta un 5,02.

Por último, cabe resaltar el hecho de que entre 2011 y 2012 se produjo un descenso de la nota media de todas las asignaturas salvo de una, que es Estadística.

6.1.5. Análisis asignaturas de tercer curso

Nota máxima, nota mínima y media

El tercer curso de Grado en Ingeniería en Tecnologías Industriales tiene un total de 60 créditos repartidos en seis asignaturas en el primer cuatrimestre, y otras seis en el segundo. Siguiendo la misma estructura que para el primer y segundo curso, continuación se muestran las notas medias, máximas, mínimas y el número de aprobados y suspensos para cada una de las asignaturas de tercero.

Asignatura	N	Nota Mínima	Nota Máxima	Media	Aprobados	Suspensos
Electrotecnia	1.158	0,0	9,7	5,03	795	363 (31,35 %)
Maquinas Eléctricas	673	1,2	10,0	6,42	619	54 (8,02 %)
Mecánica Fluidos	715	0,0	9,0	5,08	542	173 (24,2 %)
Mecánica Medios Continuos	1.032	0,0	10,0	5,50	754	278 (26,94 %)
Optimización Simulación	710	0,0	9,8	5,40	508	202 (28,45 %)
Organización Gestión	698	0,0	9,6	6,13	611	87 (12,46 %)
Proyecto II	619	5,0	10,0	8,42	619	0 (0 %)
Resistencia Materiales	667	0,0	9,2	5,93	568	99 (14,84 %)
Técnicas Estadísticas Calidad	964	0,0	10,0	6,25	906	58 (6,02 %)
Tecnología Medio Ambiente y Sostenibilidad	1.111	0,0	9,5	5,56	826	285 (25,65 %)
Tecnología Selección Materiales	1.029	0,0	9,3	5,64	817	212 (20,6 %)
Termodinámica	1.042	0,0	9,7	5,89	878	164 (15,74 %)

Tabla 6.14. Estadísticas generales de las distintas asignaturas de tercero.

En primera instancia, se puede apreciar como existen dos asignaturas cuya nota mínima entre 2012 y 2014 no es un cero. Una de ellas es máquinas eléctricas con un 1,2; y la otra es Proyecto II con un 5,0, por lo que en tres años no ha suspendido nadie esta asignatura. En cuanto a las notas máximas, sólo las de cuatro asignaturas son un 10, que son Máquinas Eléctricas, Mecánica de Medios Continuos, Proyecto II y Técnicas Estadísticas para la Calidad. El resto de asignaturas poseen una nota máxima igual o superior a 9,0. En cuanto a las notas medias de tercero, todas ellas se encuentran por encima del cinco. La nota media más alta, análogamente al segundo curso, que era Proyecto I, en este caso es Proyecto II con un valor de 8,42. Por el contrario, la nota media más baja de tercero es la de Electrotecnia con un 5,03.

La asignatura de tercero con un porcentaje mayor de suspensos es Electrotecnia con un 31,35 %, seguida de Optimización y Simulación que posee un 28,45 %. La asignatura con un porcentaje menor de suspensos es Proyecto II, que tal y como se ha dicho no se ha registrado ningún suspenso a lo largo de estos años. Máquinas Eléctricas y Técnicas Estadísticas para la Calidad también presentan un porcentaje de suspensos bajo, con unos valores de 8,02 % y de 6,02 % respectivamente.

Varianza, Asimetría y Curtosis

A continuación se observa la varianza, la asimetría y la curtosis de las asignaturas de tercero para determinar si hay algún valor que llame especialmente la atención.

Asignatura	Varianza	Asimetría	Curtosis
Electrotecnia	2,85	-0,549	0,322
Maquinas Eléctricas	1,85	-0,043	0,635
Mecánica Fluidos	1,95	-1,195	3,601
Mecánica Medios Continuos	2,47	-0,355	0,986
Optimización Simulación	2,87	-0,798	1,720
Organización Gestión	1,92	-0,746	2,265
Proyecto II	0,71	-0,756	1,584
Resistencia Materiales	1,98	-1,105	3,642
Técnicas Estadísticas Calidad	1,29	-0,704	4,186
Tecnología Medio Ambiente y Sostenibilidad	2,41	-0,455	1,409
Tecnología Selección Materiales	1,71	-0,518	1,700
Termodinámica	2,49	-0,813	2,641

Tabla 6.15. Varianza, asimetría y curtosis de las asignaturas de tercero.

Se puede apreciar como en el tercer curso las varianzas en general son más bajas, y no hay ninguna excesivamente alta que destaque sobre el resto, tal y como sucedía en primero y en segundo. Una explicación a este hecho podría ser el filtro natural que se realiza sobre los alumnos a medida que pasan los cursos. Dicho de otro modo, muchos de los alumnos que empiezan primero y obtienen resultados extremadamente malos (aumentando así la varianza) ya no llegan a segundo por no superar la fase selectiva, y una parte de los que la superan pero con malos resultados, abandona los estudios antes de llegar a tercero.

En cuanto a las asimetrías, del mismo modo que en primer y segundo curso, son todas negativas, por lo que la cola de la distribución en todas las asignaturas se encuentra a la izquierda de la media. Sobre los valores de la curtosis, se observa que el más alto es el que corresponde a la asignatura Técnicas Estadísticas para la Calidad. Realizando un histograma de las notas de esta asignatura, se puede observar como prácticamente todos los valores se encuentran en el intervalo comprendido entre 4 y 8, siendo la media de 6,25.

Evolución temporal

La media de las asignaturas de tercer curso desde el año 2012 hasta el año 2014 se muestran a continuación.

Asignatura	2012	2013	2014
Electrotecnia	5,03	4,97	5,11
Maquinas Eléctricas	6,17	6,40	6,76
Mecánica Fluidos	5,93	4,84	4,80
Mecánica Medios Continuos	6,06	5,32	5,18
Optimización Simulación	6,08	5,24	5,10
Organización Gestión	6,73	6,12	5,34
Proyecto II	8,48	8,46	8,25
Resistencia Materiales	6,83	5,87	5,08
Técnicas Estadísticas Calidad	5,94	6,21	6,72
Tecnología Medio Ambiente y Sostenibilidad	5,94	5,15	5,84
Tecnología Selección Materiales	5,98	5,24	5,89
Termodinámica	5,93	5,42	6,60

Tabla 6.16. Evolución temporal de las notas de las asignaturas de tercero.

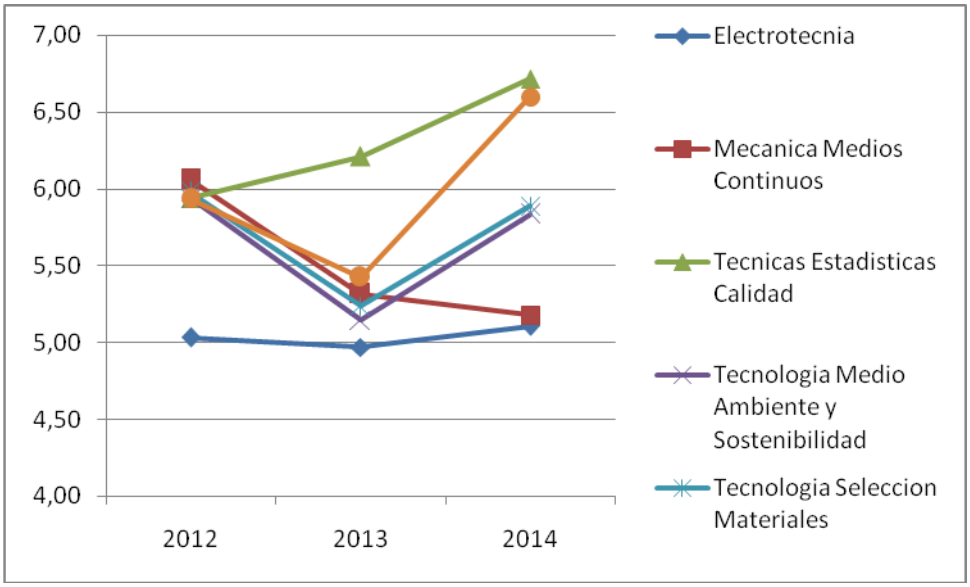

Figura 6.5. Evolución temporal de las asignaturas del primer cuatrimestre.

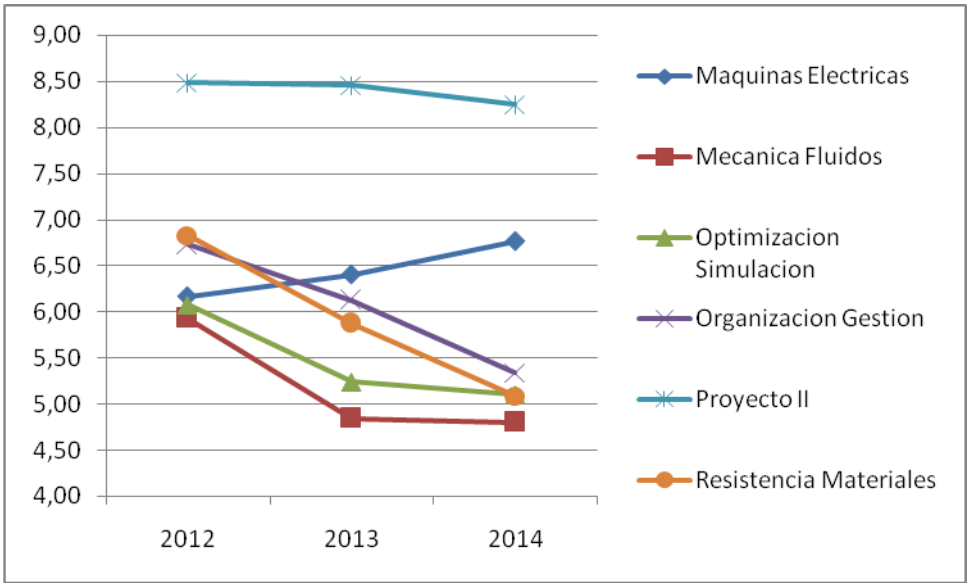

Figura 6.6. Evolución temporal de las asignaturas del segundo cuatrimestre.

Viendo las gráficas, se observa como la media de Proyecto II destaca sobre el resto durante todos los años. También se observa como la media de Electrotecnia se mantiene baja en el tiempo, y como Mecánica de Fluidos protagoniza una bajada brusca entre 2012 y 2013, situando la nueva media por debajo del cinco en los dos últimos años.

La nota media más alta registrada en estos tres años se encuentra en Proyecto II en el 2012 con un valor de 8,48, y la más baja se observa en Mecánica de Fluidos en el 2014 con un valor de 4,80.

En cuanto a las variaciones, la más grande en positivo es la sucedida en Termodinámica entre los años 2013 y 2014, protagonizando un aumento del 21,6 % (de 5,42 a 6,60), y la mayor variación negativa es la encontrada en Mecánica de Fluidos entre 2012 y 2013, cuyo valor sufre un descenso del 18,3 % (de 5,93 a 4,84).

En cuanto a las tendencias, cabe destacar el gran número de asignaturas con una tendencia negativa, que son: Mecánica de Fluidos, Mecánica de Medios Continuos, Optimización y Simulación, Organización de la Gestión, Proyecto II, Resistencia de Materiales. Todas ellas salvo Mecánica de Medios Continuos, pertenecen al segundo cuatrimestre. Por el contrario, las asignaturas que presentan una tendencia positiva son Máquinas Eléctricas y Técnicas Estadísticas para la Calidad.

En el tercer curso sucede algo parecido a lo que sucede en el segundo curso, y es que de un año para otro (de 2012 a 2013 en este caso), se experimenta un descenso de la nota media de todas las asignaturas salvo de dos, que son Máquinas Eléctricas y Técnicas Estadísticas para la calidad.

6.1.6. Análisis asignaturas de cuarto curso

Nota máxima, nota mínima y media

El último curso de Grado en Ingeniería en Tecnologías Industriales consta de 42 créditos troncales, 12 de los cuales pertenecen al Trabajo de Fin de Grado (TFG). Las notas medias, máximas, mínimas y el número de aprobados y suspensos para cada una de las asignaturas de cuarto se muestran a continuación:

Asignatura	N	Nota Mínima	Nota Máxima	Media	Aprobados	Suspensos
Control Automático	514	0,0	9,1	5,32	336	178 (34,63 %)
Electrónica	499	0,0	9,4	6,15	454	45 (9,02 %)
Gestión Proyectos	477	1,0	9,0	6,92	474	3 (0,63 %)
Sistemas Fabricación	485	0,0	9,5	7,14	469	16 (3,3 %)
Termotecnia	507	0,0	10,0	6,62	424	83 (16,37 %)
TFG	185	5,5	10,0	8,73	185	0 (0 %)

Tabla 6.17. Estadísticas generales de las distintas asignaturas de cuarto.

En el cuarto curso, las estadísticas calculadas son de los años 2013 y 2014. En dichos años, se observa que hay dos asignaturas en las que no se ha sacado un 0,0, que son Gestión de Proyectos, cuya nota mínima es un 1,0, y el TFG, cuya nota mínima es un 5,5. En cuanto a las notas máximas, todas ellas están por encima del 9,0, pudiéndose observar un diez para Termotecnia, y otro para el TFG.

Las notas medias de cuarto son más elevadas que las de cursos anteriores. Todas ellas menos una (Control Automático) son superiores al seis, siendo la del TFG la más alta con un 8,73, y Control Automático la más baja con un 5,32.

La asignatura con menor porcentaje de suspensos es el TFG, que en los dos años de los que se disponen datos no ha suspendido ningún estudiante. Le sigue de cerca Gestión de Proyectos, cuyo porcentaje de suspensos es de 0,63 %. Por el contrario, la asignatura más suspendida es Control Automático, con un 34,63 % de suspensos.

Varianza, Asimetría y Curtosis

Se muestra a continuación los valores obtenidos para cada asignatura de la varianza, la asimetría y la curtosis:

Asignatura	Varianza	Asimetría	Curtosis
Control Automático	3,34	-0,585	0,038
Electrónica	1,93	-1,082	4,091
Gestión Proyectos	0,60	-1,035	6,838
Sistemas Fabricación	1,73	-1,553	6,052
Termotecnia	4,30	-1,244	2,110
TFG	1,04	-0,944	0,455

Tabla 6.18. Varianza, asimetría y curtosis de las asignaturas de cuarto

Se observa una varianza especialmente alta para la asignatura de Termotecnia. No obstante, el valor de la curtosis de 2,110, el cual el ser superior a cero, indica que la distribución es leptocúrtica, por lo que los datos se concentran alrededor de la región central de la distribución. En capítulos posteriores se realizará un histograma de las notas obtenidas por los estudiantes en dicha asignatura para ver qué sucede.

Del mismo modo que en cursos anteriores, la cola de la distribución para todas las asignaturas se encuentra a la izquierda de la media, que es lo que indica un valor negativo de la asimetría.

En cuanto a la curtósis, se observan dos valores por encima del resto, que son los correspondientes a Gestión de Proyectos y a Sistemas de Fabricación. Dado que el valor de su varianza es bajo, se puede afirmar que las notas de dichas asignaturas están muy concentradas alrededor de la media.

Evolución temporal

Por último, es interesante mostrar las variaciones producidas en las notas de las asignaturas de cuarto en los dos años que se poseen datos.

Asignatura	2013	2014
Control Automático	6,00	4,55
Electrónica	6,15	6,16
Gestión Proyectos	7,00	6,82
Sistemas Fabricación	7,01	7,30
Termotecnia	6,84	6,35
TFG	8,70	8,85

Tabla 6.19. Evolución temporal de las notas de cuarto.

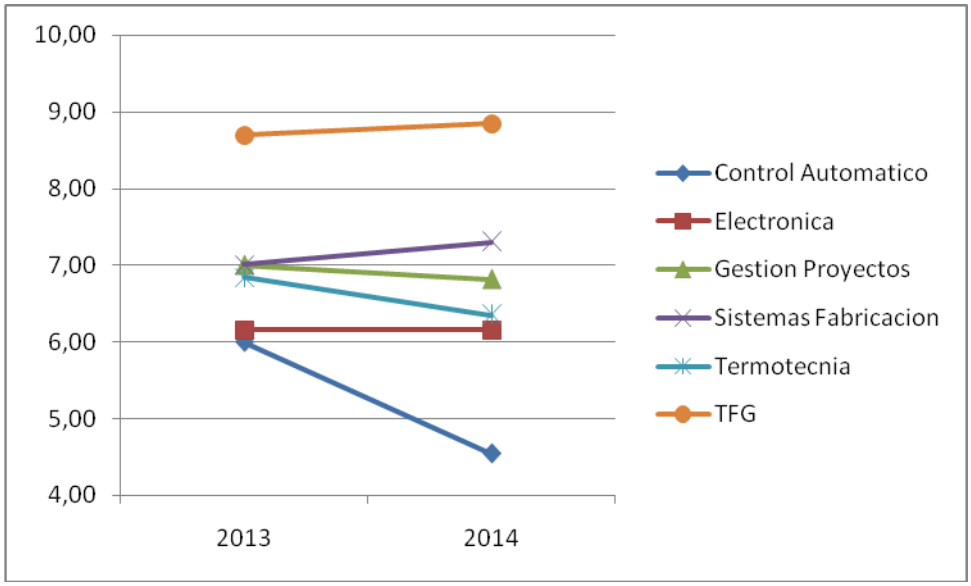

Figura 6.7. Evolución temporal de las notas de cuarto.

Se observa que la media del TFG destaca por encima del resto tanto en 2013 como en 2014. También llama la atención el descenso del 24 % de la nota media de Control

Automático, pasando de un 6,00 a un 4,55. La media de Electrónica se mantiene constante de un año para el otro, y el resto de asignaturas experimentan subidas o bajadas no especialmente grandes.

6.1.7. Análisis de asignaturas que presentan una distribución peculiar

Fundamentos de Informática

Fundamentos de Informática es una de las asignaturas que corresponde al primer cuatrimestre del primer curso que ha llamado la atención por sus valores. Es la asignatura con un mayor porcentaje de suspensos en primero, y destaca su alta varianza y el valor negativo de la curtosis, que indica que los datos están poco concentrados en la región central de la distribución.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Fundamentos Informática	2.472	5,28	834 (33,74 %)	6,52	-0,482	-0,588

Se muestra a continuación un histograma con las notas finales obtenidas en dicha asignatura con el objetivo de ver qué forma tiene la distribución:

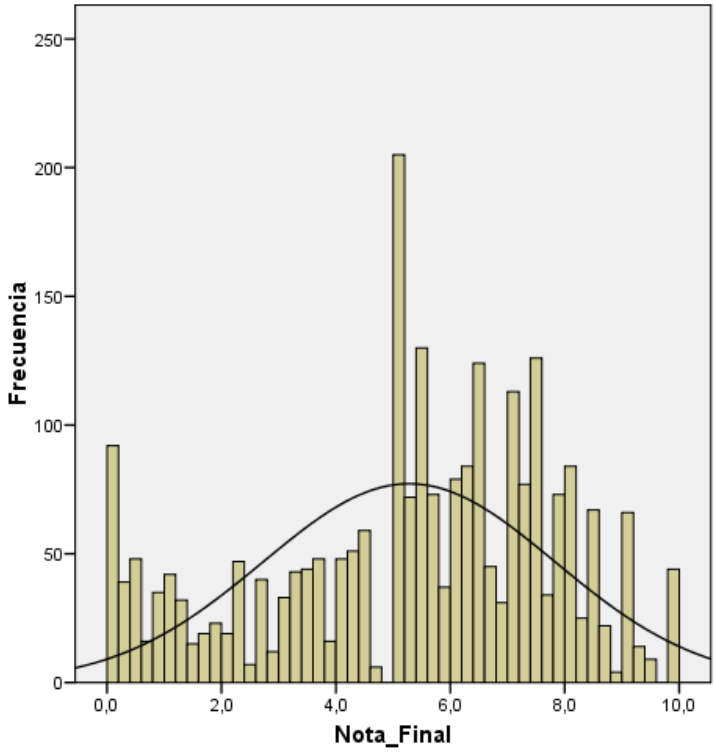

Figura 6.8. Histograma de las notas de Fundamentos de Informática

En primera instancia, se observa como la distribución no se ajusta a una distribución normal, que se ha representado sobre el gráfico con una línea negra. Existen numerosos valores superiores a la media (5,28) por encima de la distribución normal, y numerosos valores extremadamente bajos (inferiores a 1,40) por debajo de dicha distribución. Hay 83 notas de 0,0, que es un 3,3 % de la muestra total, y 304 valores inferiores al 1,4, representando un 12,3 % de la muestra total. Destaca también el hecho de que no haya ninguna nota igual a 4,8 ni 4,9, y el 5,0 sea el valor con mayor frecuencia (un 7,0 % del total), por lo que se deduce que se ha realizado un arrastre de una o dos décimas hasta llegar al 5,0. Este hecho sucede también en menor escala para todos aquellos valores en que cambia la cualificación del alumno (de aprobado a notable, y de notable a excelente). Si se observan los valores que delimitan dicho cambio de cualificación (el 7,0 y el 9,0), se observa como el intervalo de notas anterior tiene una frecuencia muy baja, mientras que el siguiente tiene una frecuencia muy elevada).

Si se calcula el valor de la media para los registros suspendidos, y el valor de la media para los registros aprobados, se puede aprecia una gran diferencia:

	Nota media
Aprobados	6,81
Suspendidos	2,28

Se concluye que la manera de evaluar dicha asignatura es muy polarizada, pues los resultados obtenidos son muy extremos. Es muy probable que los ejercicios de examen sean del tipo "o se sabe hacer, o no se sabe", motivo por el cual las notas son o muy bajas, o muy altas.

Puede resultar interesante calcular el número de veces que un alumno cursa en promedio Fundamentos de Informática. Para evitar distorsiones en el resultado, se ha realizado el cálculo únicamente con aquellos estudiantes que finalmente han aprobado la asignatura.

- Nº Alumnos = 1638
- Nº Aprobados = 1638
- Nº Suspensos = 597
- $\Pr omedio\_FundamentosInformatica = \frac{1638 + 597}{1638} = 1{,}36$

En promedio, un alumno cursa Fundamentos de Informática 1,36 veces.

Expresión Gráfica

La asignatura del segundo cuatrimestre del primer curso Expresión Gráfica también destaca por tener un valor de la varianza alto. Dicho valor es menor que en el caso de Fundamentos de Informática, pero es superior al resto de asignaturas.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Expresión Gráfica	2.029	5,51	449 (22,13 %)	4,30	-0,806	0,661

Procediendo del mismo modo que con Fundamentos de Informática, se realiza un histograma para observar la forma de su distribución:

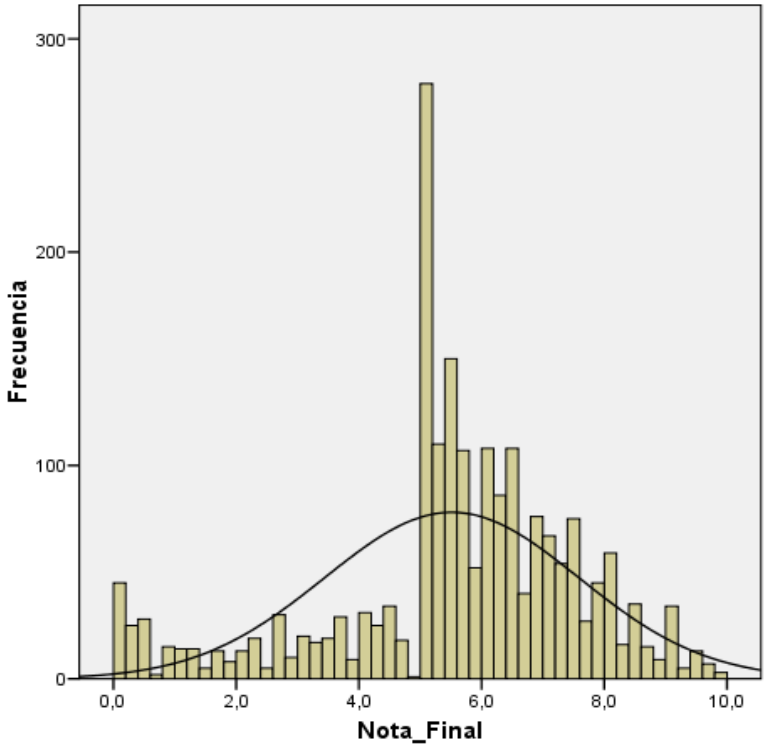

Figura 6.9. Histograma de las notas de Expresión Gráfica

Se observan dos partes muy diferenciadas. Las notas iguales o superiores del 5,0, que son el 77,87 %, se distribuyen de una forma muy similar a una distribución normal pero con unos valores algo más altos. Sin embargo, las notas inferiores al 5,0 presentan una distribución bastante constante. De las 22,13 % notas suspendidas, aproximadamente un 50 % se encuentran entre 0 y 2,5; y el otro 50 % se encuentra entre 2,5 y 4,9. Se puede comprobar cómo en este caso, se realiza también un arrastre de las notas de 4,8 y 4,9 hacia el 5,0.

La nota promedio de los suspensos y de los aprobados es la siguiente:

	Nota media
Aprobados	6,38
Suspendidos	2,46

En Expresión Gráfica, se observa como también existe una gran diferencia entre la nota media de los suspensos y la de los aprobados. En cuanto a los suspensos no se puede extraer ninguna conclusión clara. En cuanto a los aprobados, se puede intuir que la manera de evaluar está muy pautada por niveles que van subiendo de dificultad progresivamente, de manera que a más nivel, cada vez hay menos alumnos que lo resuelven con éxito.

El promedio de veces que un alumno cursa Expresión Gráfica es el siguiente:

- Nº Alumnos = 1579
- Nº Aprobados = 1579
- Nº Suspensos = 301
- $\mathrm{Pr}\,omedio\_ExpresionGrafica = \frac{1579+301}{1579} = 1{,}19$

Informática

La asignatura del primer cuatrimestre del segundo curso Informática destaca por tener un valor de la varianza alto con respecto al resto de asignaturas del mismo curso.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Informática	1.573	5,84	322 (20,47 %)	3,84	-0,646	0,794

Si se realiza un histograma con las notas registradas en la asignatura se obtiene la siguiente distribución:

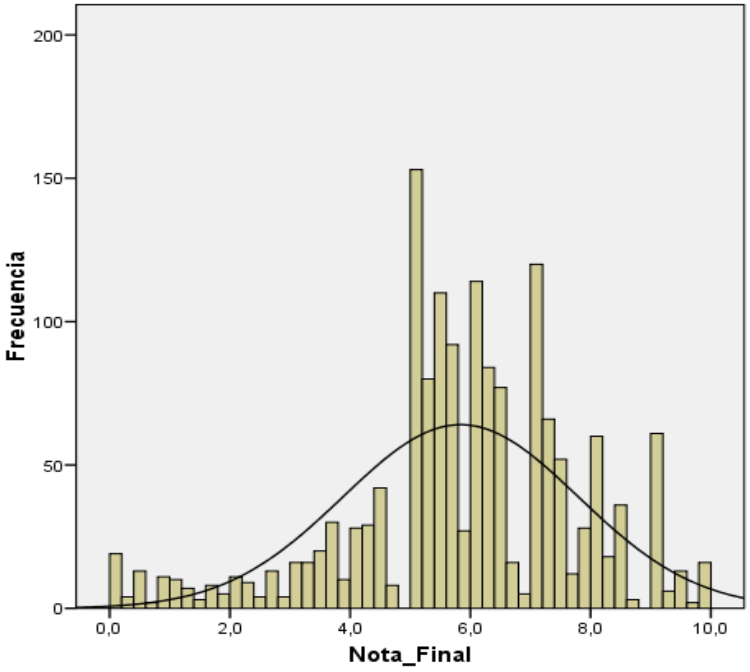

Figura 6.10. Histograma de las notas de Informática

Se observa una distribución cuyos valores por encima del 5,0 tienen una frecuencia mayor a los que marca la distribución normal, y lo que están por debajo tienen una frecuencia menor a la marcada por la normal, salvo los casos más extremos, que están por encima. Del mismo modo que sucede con el resto de asignaturas analizadas, se suben las notas de 4,8 y 4,9 a 5,0. Se observa también cómo existen picos en el histograma, que a parte del 5,0, corresponden a las notas superiores a 5,0 con un valor redondo (6,0; 7,0; 8,0; 9,0 y 10,0). Dichas notas tienen una frecuencia mucho mayor a su intervalo de notas inmediatamente inferior, por lo que se deduce que se produce un arrastre también. En la asignatura Fundamentos de Informática sucedía el mismo hecho.

La media de las notas medias aprobadas y suspendidas son las siguientes:

	Nota media
Aprobados	6,58
Suspendidos	2,94

A diferencia de Fundamentos de Informática, las notas no están tan polarizadas, ya que las notas medias de los aprobados y de los suspendidos no distan tanto como en Fundamentos de Informática.

El promedio de veces que un alumno cursa informática es el siguiente:

- Nº Alumnos = 1251
- Nº Aprobados = 1251
- Nº Suspensos = 177
- $\text{Promedio\_Informatica} = \frac{1251+177}{1251} = 1{,}14$

Mecánica

Mecánica es una asignatura del primer cuatrimestre del segundo curso. Es la asignatura con una nota media más baja de toda la carrera y la que tiene un mayor porcentaje de suspensos, cuyo valor es cercano al 50 %.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Mecánica	2.058	4,33	1003 (48,74 %)	3,47	-0,306	0,552

Se realiza un histograma con las notas de la asignatura para observar cómo se distribuyen:

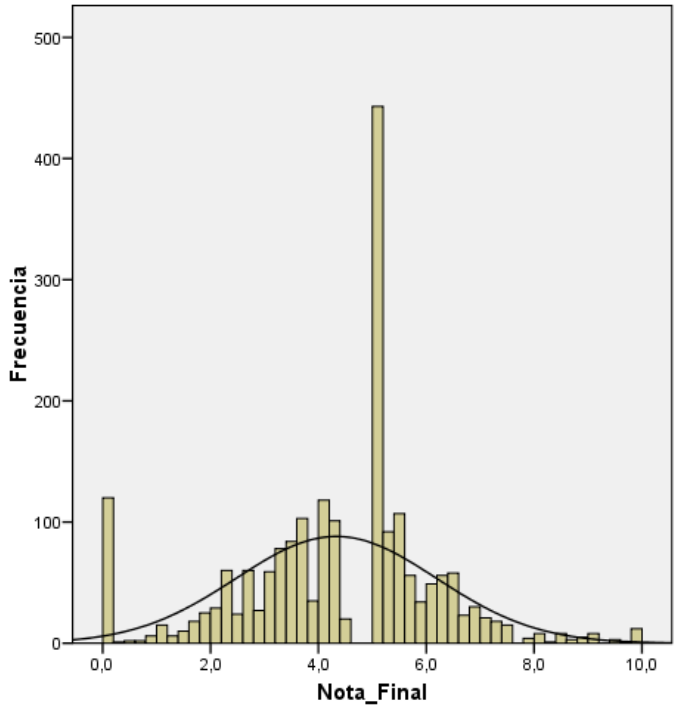

Figura 6.11. Histograma de las notas de Mecánica

Se puede observar cómo, quitando un par de picos que sobresalen sobre el resto (los situados en 0,0 y 5,0), y una franja sin notas (de 4,5 a 4,9), la campana se ajusta bastante a una distribución normal con la media situada alrededor del 4,3. Por un lado, se observa una cantidad exagerada de 0,0, concretamente 120, lo que representa un 5,8 % del total de la muestra. Por otro lado, 382 notas son un 5,0, lo que representa un 18,6 % de los datos totales. Observado una tabla de frecuencias, se comprueba que, debido a las bajas notas en la asignatura, el arrastre de notas hasta el 5,0 se realiza desde 4,5. Es por este motivo que la frecuencia en este punto es tan elevada.

Si calculamos la media de la asignatura de las notas por encima y por debajo de 5,0, el resultado es el siguiente:

	Nota media
Aprobados	5,74
Suspendidos	2,85

Mientras que la media de los suspendidos se mantiene en un valor parecido al del resto de asignaturas, la media de aprobados es notablemente más baja que el resto, ya que está alrededor de un punto por debajo.

Con los estudiantes que finalmente han aprobado la asignatura (hayan repetido o no), se calcula el número de veces que en promedio cursa Mecánica:

- Nº Alumnos = 1054
- Nº Aprobados = 1054
- Nº Suspensos = 417
- $\Pr omedio\_Mecanica = \frac{1054 + 417}{1054} = 1{,}40$

En promedio, un estudiante cursa Mecánica 1,48 veces.

Proyecto I

En el segundo cuatrimestre del segundo curso se imparte la asignatura Proyecto I, la cual destaca por su elevada media, su bajo número de suspensos, y por tener una curtosis excesivamente alta, hecho que indica que los datos están muy centralizados alrededor de la media.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Proyecto I	1.087	8,37	2 (0,18 %)	1,07	-1,245	7,061

Se representan las notas de la asignatura en un histograma:

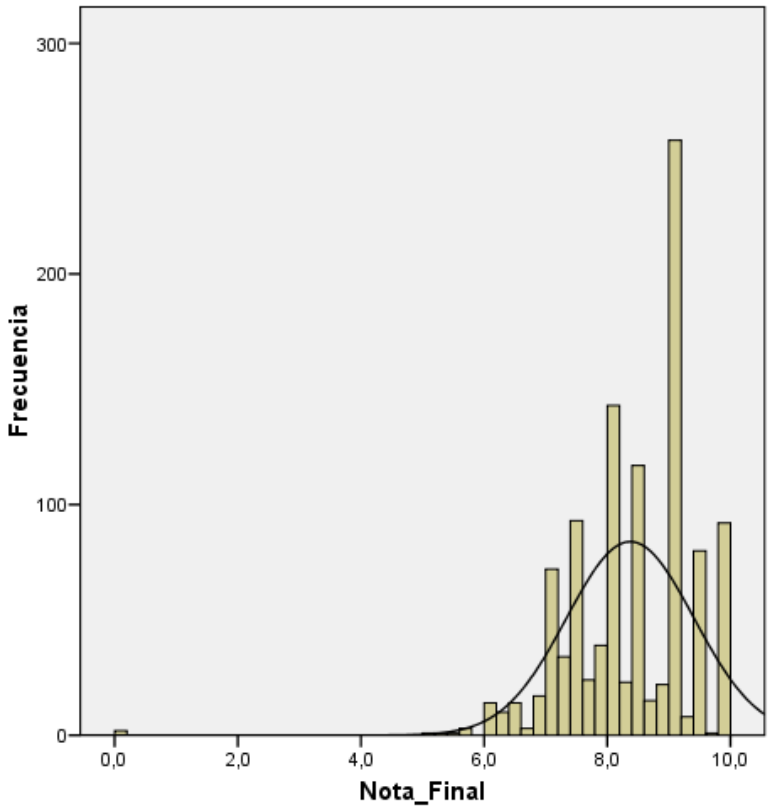

Figura 6.12. Histograma de las notas de Proyecto I

Tal y como se sospechaba, los datos están muy concentrados en una región de la distribución alrededor de la media. La nota más obtenida es un 9,0, que representa el 20,9 % de los casos. Sorprende observar cómo más del 99,9 % de los casos están comprendidos entre un 6,0 y un 10,0.

Para Proyecto I, la nota media de los suspendidos es un 0,0 debido a dos casos aislados, y la nota media de los aprobados es un 8,39.

Dado que únicamente hay dos suspensos, el promedio de veces que se cursa Proyecto I es 1,00.

Técnicas Estadísticas para la Calidad

Técnicas Estadísticas para la Calidad es una asignatura impartida en el primer cuatrimestre del tercer curso. Dicha asignatura es analizada a parte por presentar un valor de la curtosis mayor que el resto de asignaturas.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Técnicas Estadísticas Calidad	964	6,25	58 (6,02 %)	1,29	-0,704	4,186

A continuación se muestra un histograma con la distribución de notas en la asignatura:

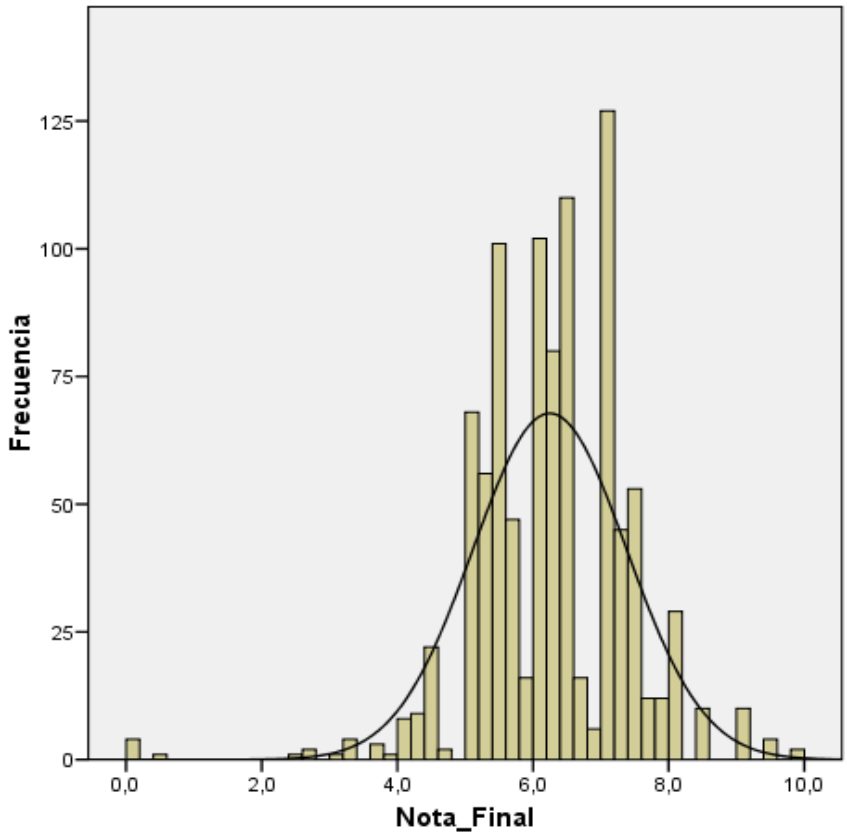

Figura 6.13. Histograma de las notas de Técnicas Estadísticas para la Calidad

El alto valor de la curtosis se debe a que el 95 % de las notas están comprendidas entre 4,0 y 8,0, siendo la media de 6,25. Se observa que se produce también un arrastre de las notas en que hay cambio de cualificación. No obstante, en este caso el valor con más frecuencia

no es el 5,0, sino que el mayor arrastre se produce de 6,9 a 7,0, siendo este el pico que más sobresale en el histograma.

Si se calcula la media de las notas suspendidas y de las aprobadas, se obtienen los siguientes resultados:

	Nota media
Aprobados	6,41
Suspendidos	3,77

En este caso, la nota media de los aprobados se mantiene cercana a la del resto de asignaturas. Sin embargo, se aprecia una media especialmente alta para las notas suspendidas, ya que está prácticamente un punto por encima de la del resto de asignaturas.

Técnicas Estadísticas para la Calidad no es una asignatura que se caracterice por tener un número de suspensos especialmente alto, pero aún así se calcula el número de veces que el estudiante la cursa en promedio:

- Nº Alumnos = 906
- Nº Aprobados = 906
- Nº Suspensos = 42
- $\Pr omedio\_Tec.EstadisticasCal = \frac{906+42}{906} = 1{,}05$

Se observa como los alumnos prácticamente cursan la asignatura una única vez, ya que apenas hay suspensos.

Termotecnia

Termotecnia es una asignatura impartida en el primer cuatrimestre del cuarto y último curso. Se ha incluido en este apartado debido a que tiene una varianza especialmente alta.

Asignatura	N	Media	Suspensos	Varianza	Asimetría	Curtosis
Termotecnia	507	6,62	83 (16,37 %)	4,30	-1,244	2,110

Se muestra el histograma de sus notas para ver qué forma tiene su distribución:

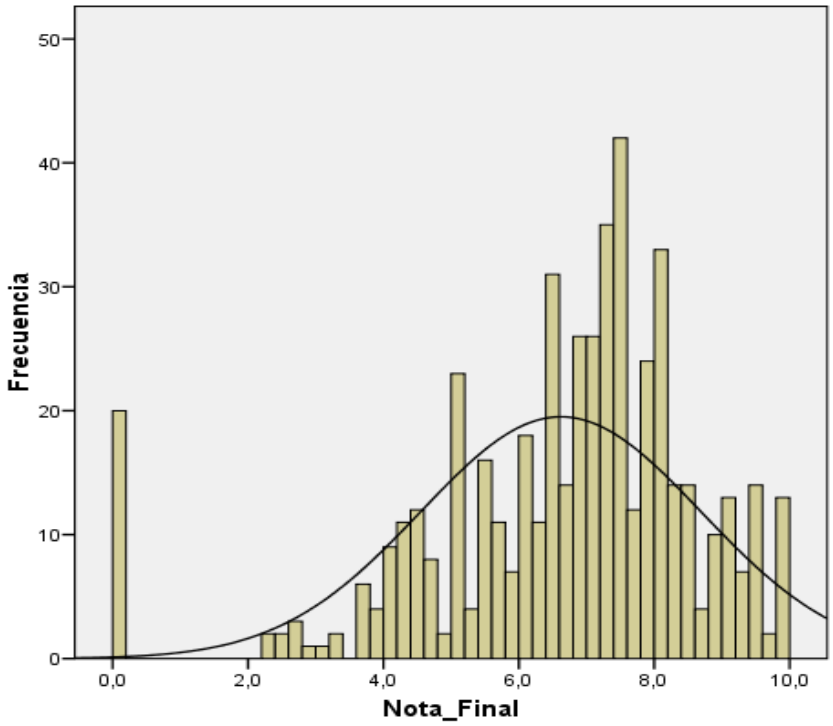

Figura 6.14. Histograma de las notas de Termotecnia

Observando el histograma, se encuentra la explicación al valor tan alto de la varianza, y es que hay una gran cantidad de 0,0. Termotecnia no es una de las asignaturas que más se suspenden (16,37 % de suspensos), pero del total de suspensos, un 24,1 % son con un 0,0. Esto se puede deber a una alta tasa de abandono en la asignatura, o a un determinado perfil de estudiantes que se les da extremadamente mal dicha asignatura. En cuanto al arrastre de notas, no existe una evidencia clara de que se realicen.

Se calcula a continuación la media de los suspensos y la media de los aprobados:

	Nota media
Aprobados	7,32
Suspendidos	3,05

Pese a la gran cantidad de 0,0 que hay, la media de los suspensos no es extremadamente baja. Sin embargo, sí que destacada la media de los aprobados por ser elevada en comparación con otras asignaturas (quitando excepciones como Proyecto I o el TFG).

Como en el resto de asignaturas analizadas, se calcula ahora el promedio de veces que los estudiantes cursan la asignatura:

- Nº Alumnos = 424
- Nº Aprobados = 424
- Nº Suspensos = 21
- $\Pr omedio\_Termotecnia = \frac{424+21}{424} = 1{,}06$

Como ya se ha dicho, Termotecnia no es una asignatura de las que se repiten especialmente, por lo que en promedio el estudiante la cursa 1,06 veces.

6.2. Estudio por sexos

6.2.1. Variables cuantitativas.

En este apartado se determinará si existen diferencias significativas en distintas variables entre hombres y mujeres. El objetivo de este análisis es saber si el sexo es una variable que permita discernir entre los resultados cuantitativos obtenidos en primero. Para saber si se pueden aplicar pruebas paramétricas, se aplica en primer lugar la prueba de Kolmogorov - Smirnov para comprobar el supuesto de normalidad, y el test de Levene para comprobar el supuesto de homocedasticidad. Se obtienen los siguientes resultados:

Kolmogorov - Smirnov

Se aplica el test a las variables cuantitativas mencionadas en el párrafo anterior para las mujeres (SexoCodif = 0) y para los hombres (SexoCodif = 1). El tamaño de la muestra es de N = 403 para las mujeres y de N'= 1408 para los hombres. Los resultados son los siguientes:

Prueba de Kolmogorov-Smirnov para una muestra[c]

		Promedio_ Primero	Alpha_ Primero	Repetidas_ Primero	Nota_Sele
N		403	403	403	403
Parámetros normales [a,b]	Media	6,0579	,7811	2,94	11,3012
	Desviación típica	1,10479	,24517	3,409	1,02219
Diferencias más extremas	Absoluta	,165	,186	,194	,050
	Positiva	,084	,186	,177	,023
	Negativa	-,165	-,181	-,194	-,050
Z de Kolmogorov-Smirnov		3,305	3,734	3,898	,998
Sig. asintót. (bilateral)		,000	,000	,000	,272

a. La distribución de contraste es la Normal.
b. Se han calculado a partir de los datos.
c. SexoCod = 0

Tabla 6.20. Prueba de Kolmogorov Smirnov para las mujeres.

Prueba de Kolmogorov-Smirnov para una muestra

		Promedio_ Primero	Alpha_ Primero	Repetidas_ Primero	Nota_Sele
N		1408	1408	1408	1408
Parámetros normales [a,b]	Media	5,9396	,7450	3,44	11,0146
	Desviación típica	1,34106	,27641	3,881	1,08670
Diferencias más extremas	Absoluta	,181	,178	,187	,065
	Positiva	,074	,178	,165	,025
	Negativa	-,181	-,166	-,187	-,065
Z de Kolmogorov-Smirnov		6,799	6,685	7,031	2,448
Sig. asintót. (bilateral)		,000	,000	,000	,000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

c. SexoCod = 1

Tabla 6.21. Prueba de Kolmogorov Smirnov para los hombres.

Se puede observar que la sigma asintótica (bilateral) es inferior al nivel de significación de 0,05 en todas las variables salvo en la "Nota_Sele" en el grupo de las mujeres (SexoCodif = 0). Esto significa que únicamente esta última variable se puede aproximar por una distribución normal. Dado que el resto de variables no satisfacen el supuesto de normalidad, se descarta el uso de cualquier prueba paramétrica para estas variables.

Test de Levene

Aunque se ha descartado el uso de pruebas paramétricas para todas las variables salvo "Nota_Sele", se ha aplicado también el test de Levene para observar si las variables cumplen el supuesto de homocedasticidad. Los resultados obtenidos son los siguientes:

		Prueba de Levene para la igualdad de varianzas	
		F	Sig.
Nota_Sele	Se han asumido varianzas iguales	,504	,478
	No se han asumido varianzas iguales		
Promedio_Primero	Se han asumido varianzas iguales	7,702	,006
	No se han asumido varianzas iguales		
Repetidas_Primero	Se han asumido varianzas iguales	10,688	,001
	No se han asumido varianzas iguales		
Alpha_Primero	Se han asumido varianzas iguales	7,424	,006
	No se han asumido varianzas iguales		

Tabla 6.22. Resultados del test de Levene

Se observa que la sigma es inferior a 0,05 en todas las variables salvo en la "Nota_Sele". Por lo tanto, se puede afirmar que hay igualdad de varianzas en los grupos en todas las variables menos en la anteriormente mencionada. Debido a que la variable "Nota_Sele" (que era la única que cumplía el supuesto de normalidad) incumple el supuesto de homocedasticidad, se descarta la aplicación de cualquier prueba paramétrica sobre dicha variable.

Dado que ninguna de las variables cumple con los dos requisitos para la aplicación de pruebas paramétricas, se recurre al test no paramétrico de la U de Mann-Whitney para muestras independientes.

Test U de Mann-Whitney

Consiste en la alternativa no paramétrica de la prueba t-student. En este caso se trata de observar si existe una relación entre el sexo del estudiante y las variables "Promedio_Primero", "Alpha_Primero" y "Repetidas_Primero". Los resultados obtenidos son los siguientes:

Rangos

	SexoCod	N	Rango promedio	Suma de rangos
Promedio_Primero	0	403	924,04	372386,50
	1	1408	900,84	1268380
	Total	1811		
Repetidas_Primero	0	403	864,73	348485,00
	1	1408	917,81	1292281
	Total	1811		
Alpha_Primero	0	403	947,68	381913,50
	1	1408	894,07	1258853
	Total	1811		

Tabla 6.23. Rangos calculados en la prueba de la U de Mann

Estadísticos de contraste[a]

	Promedio_ Primero	Repetidas_ Primero	Alpha_ Primero
U de Mann-Whitney	276443,500	267079,000	266916,500
W de Wilcoxon	1268379,500	348485,000	1258852,500
Z	-,785	-1,840	-1,855
Sig. asintót. (bilateral)	,432	,066	,064

a. Variable de agrupación: SexoCod

Tabla 6.24. Sigma asintótica de la prueba de la U de Mann Whitney.

Se observa una sigma asintótica mayor que el nivel de significación de 0,05, por lo que se acepta la hipótesis nula de que los datos proceden de la misma población. Este hecho se traduce en que no existen diferencias significativas en las variables estudiadas entre hombres y mujeres, por lo que se concluye que el sexo no es un buen predictor para el promedio de primero y para las asignaturas repetidas.

6.2.2. Variables cualitativas.

En este apartado se desea conocer si el sexo es una buena variable para diferenciar entre los grupos creados para los estudiantes de primero. Recordemos que las variables cualitativas que contienen dichos grupos son: Grupos_Primero (estudiantes por encima de la media de primero indicados con un 1, y estudiantes por debajo de la media de primero

marcados con un 0); Repiten_Primero_S_N (el grupo 0 no repite ninguna asignatura, y el grupo 1 repite una o más asignaturas); y por último la variable Completan_Primero_Al_Ritmo (toma el valor 0 si el alumno no completa primero en un año; 1 si aprueba todos los créditos de primero en un año; y 3 si ese registro no se incluye en el análisis por ser un caso excepcional).

Para observar si existe algún tipo de relación entre el sexo y las variables mencionadas en el párrafo anterior, se crea una tabla de contingencia para cada una de las variables y se aplica el estadístico Chi - Cuadrado. Los resultados obtenidos son los siguientes:

Para la variable Grupos_Primero se analizan únicamente aquellos estudiantes que han completado primero (N = 1439):

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
SexoCod * Grupos_Primero	1439	100,0%	0	,0%	1439	100,0%

Tabla de contingencia SexoCod * Grupos_Primero

			Grupos_Primero		Total
			0	1	
SexoCod	0	Recuento	218	125	343
		Frecuencia esperada	207,4	135,6	343,0
	1	Recuento	652	444	1096
		Frecuencia esperada	662,6	433,4	1096,0
Total		Recuento	870	569	1439
		Frecuencia esperada	870,0	569,0	1439,0

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	1,808[b]	1	,179		
Corrección por continuidad[a]	1,642	1	,200		
Razón de verosimilitudes	1,821	1	,177		
Estadístico exacto de Fisher				,184	,100
Asociación lineal por lineal	1,807	1	,179		
N de casos válidos	1439				

Tabla 6.25. Tabla completa del análisis de Chi - Cuadrado para las variables SexoCodif vs Grupos_Primero

En tabla anterior se puede observar en primer lugar el número de casos válidos para el análisis, una comparativa de las frecuencias observadas y las calculadas, y por último y más importante para este estudio, el parámetro sigma asintótica para la prueba de Chi - Cuadrado de Pearson, que en este caso tiene un valor de 0,179. Dado que este valor está por encima del valor de significación del 0,05, se asume que no existen diferencias en las proporciones de hombres y mujeres que están por encima o por debajo de la media total de primero.

Para las otras dos variables, Repiten_Primero_S_N y Completan_Primero _Al_Ritmo, se ha realizado exactamente el mismo estudio, y se puede encontrar completo en el Anexo. A fin de sintetizar un poco la información en esta memoria, se muestra únicamente la tabla con el parámetro sigma asintótica para la prueba de Chi - Cuadrado de Pearson. En el caso de la variable Repiten_Primero_S_N, se realiza un análisis de 1439 estudiantes, ya que únicamente se seleccionan aquellos que han completado primero. Para la variable Completan_Primero_S_N se analizan los 1797 estudiantes de la base de datos (se excluyen 14 del total por presentar datos anómalos).

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	,006[b]	1	,939		
Corrección por continuidad[a]	,000	1	,989		
Razón de verosimilitudes	,006	1	,939		
Estadístico exacto de Fisher				,950	,494
Asociación lineal por lineal	,006	1	,939		
N de casos válidos	1439				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 140,39.

Tabla 6.26. Prueba de Chi Cuadrado para las variables Sexo_Codif vs Repiten_Primero_S_N

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	2,210[b]	1	,137		
Corrección por continuidad[a]	2,041	1	,153		
Razón de verosimilitudes	2,196	1	,138		
Estadístico exacto de Fisher				,146	,077
Asociación lineal por lineal	2,209	1	,137		
N de casos válidos	1797				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 156,20.

Tabla 6.27. Prueba de Chi Cuadrado para las variables Sexo_Codif vs Completan_Primero_Al_Ritmo

Los resultados de la prueba de Chi - Cuadrado indican que no existen diferencias en la proporción de hombres y mujeres que repiten en primero ni tampoco las existen en la proporción de hombres y mujeres que completan primero o no en el primer año. Por lo tanto, el sexo no es una variable a tener en cuenta para prever el comportamiento de los estudiantes en primero.

6.3. Estudio por ubicación

6.3.1. Variables cuantitativas.

En este apartado se realizará un análisis con el objetivo de determinar si existen diferencias significativas entre los alumnos que residen en distintos emplazamientos, y entre los que han estudiado en distintas provincias o comunidades. Para llevar a cabo el estudio de las variables cuantitativas se ha realizado la prueba de Kruskal Wallis para k muestras independientes. Dicha prueba se puede considerar una extensión de la U de Mann - Whitney para 3 o más grupos. Para realizar el test, en primer lugar es necesario codificar las ubicaciones de los alumnos de la manera que se indica en la siguiente tabla:

UbicacionCodif	Ubicación	N
1	Barcelona	1436

2	Tarragona	122
3	Lleida	64
4	Girona	78
5	Islas Baleares	75
6	Otros	45
Total general		**1820**

Tabla 6.28. Codificación de las distintas ubicaciones.

Las variables analizadas son Nota_sele, Promedio_Primero, Alpha_Primero y Repetidas_Primero, y el análisis se hace para un universo de 1439 estudiantes. El resultado del test de Kruskal Wallis para los distintos lugares de residencia (UbicacionCodif) es el siguiente:

Rangos

	UbicacionCodif	N	Rango promedio
Nota_Sele	1	1140	710,97
	2	103	774,62
	3	49	824,00
	4	62	748,19
	5	54	652,56
	6	31	767,48
	Total	1439	
Promedio_Primero	1	1140	702,55
	2	103	740,02
	3	49	794,16
	4	62	824,84
	5	54	846,05
	6	31	748,89
	Total	1439	
Repetidas_Primero	1	1140	737,50
	2	103	681,88
	3	49	650,93
	4	62	611,94
	5	54	627,26
	6	31	690,06
	Total	1439	
Alpha_Primero	1	1140	702,20
	2	103	757,20
	3	49	784,28
	4	62	833,99
	5	54	812,56
	6	31	760,32
	Total	1439	

Tabla 6.29. Rangos calculados en la prueba de Kruskal Wallis.

Estadísticos de contraste[a,b]

	Nota_Sele	Promedio_ Primero	Repetidas_ Primero	Alpha_ Primero
Chi-cuadrado	7,500	12,876	12,175	12,612
gl	5	5	5	5
Sig. asintót.	,186	,025	,032	,027

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: UbicacionCodif

Tabla 6.30. Resultados de la prueba de Kruskal Wallis para los lugares de residencia del estudiante

Se puede observar como la sigma asintótica es inferior al nivel de significación 0,05 en todas las variables salvo en la "Nota_Sele". Por lo tanto, salvo para dicha variable, se rechaza la hipótesis nula de que todos los datos proceden de la misma población, hecho que conlleva que hayan diferencias significativas en los resultados obtenidos en primero entre los estudiantes de los distintos lugares de residencia.

De manera análoga al análisis por lugar de residencia, se ha repetido el proceso para el emplazamiento de la escuela donde el estudiante a cursado sus estudios. El cálculo de los rangos se puede observar en el Anexo, y la tabla con los resultados es la siguiente:

Estadísticos de contraste[a,b]

	Nota_Sele	Promedio_ Primero	Repetidas_ Primero	Alpha_ Primero
Chi-cuadrado	7,550	8,288	9,860	10,519
gl	4	4	4	4
Sig. asintót.	,110	,082	,043	,033

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Ubicacion_Escuela_Cod

Tabla 6.31. Resultados de la prueba de Kruskal Wallis para los emplazamientos de las escuelas donde ha estudiado el alumno.

En este caso, se observa una sigma asintótica inferior al nivel de significación de 0,05 en dos de las cuatro variables: "Repetidas_Primero" y "Alpha_Primero". Esto significa que el lugar

donde ha estudiado el alumno es un parámetro relevante a la hora de determinar el número de asignaturas que repetirá en primero y su parámetro Alpha (es lógico pensar que si es relevante en una de las variables, también lo será en la otra, ya que ambas están relacionadas).

6.3.2. Variables cualitativas.

De igual modo que sea ha realizado con la variable sexo, se desea comprobar en este caso si existen diferencias por la ubicación (ubicación del alumno y ubicación de la escuela) entre aquellos que están por encima o por debajo de la media (Grupos_Primero) y entre aquellos que completan primero o no (Completan_Primero_Al_Ritmo). Para ello se ha construido una tabla de contingencia y se ha aplicado el estadístico de Chi - Cuadrado.

Los resultados de la prueba de Chi - Cuadrado obtenidos para las variables UbicacionCodif (ubicación del alumno) y Grupos_Primero (alumnos por encima o por debajo de la media en primero) son los siguientes:

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
UbicacionCodif * Grupos_Primero	1439	100,0%	0	,0%	1439	100,0%

Tabla 6.32. Tabla resumen de la muestra analizada para la prueba de Chi - Cuadrado

Tabla de contingencia UbicacionCodif * Grupos_Primero

			Grupos_Primero		
			0	1	Total
UbicacionCodif	1	Recuento	717	423	1140
		Frecuencia esperada	689,2	450,8	1140,0
	2	Recuento	62	41	103
		Frecuencia esperada	62,3	40,7	103,0
	3	Recuento	26	23	49
		Frecuencia esperada	29,6	19,4	49,0
	4	Recuento	26	36	62
		Frecuencia esperada	37,5	24,5	62,0
	5	Recuento	23	31	54
		Frecuencia esperada	32,6	21,4	54,0
	6	Recuento	16	15	31
		Frecuencia esperada	18,7	12,3	31,0
Total		Recuento	870	569	1439
		Frecuencia esperada	870,0	569,0	1439,0

Tabla 6.33. Tabla de Frecuencias observadas y frecuencias esperadas.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	21,078[a]	5	,001
Razón de verosimilitudes	20,594	5	,001
Asociación lineal por lineal	17,720	1	,000
N de casos válidos	1439		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 12,26.

Tabla 6.34. Sigma asintótica para la prueba de Chi - Cuadrado.

Como se puede observar, la sigma asintótica para la prueba de Chi - Cuadrado de Pearson es de 0,001. Dado que este valor es inferior al nivel de significación marcado de 0,05, se puede afirmar que la ubicación del alumno es un factor a tener en cuenta a la hora de predecir si éste estará por encima o por debajo de la media en primero.

Para determinar si el lugar de residencia del estudiante influye a la hora de determinar si un alumno completará primero o no en el primer año, se ha vuelto a aplicar la prueba de Chi -

Cuadrado, pero en este caso sobre la variable Completan_Primero_Al_Ritmo. Esta prueba se ha aplicado sobre 1795 alumnos de la base de datos. A continuación se muestra la tabla que contiene el parámetro sigma asintótica, que es el más determinante, y el análisis completo se puede encontrar en el Anexo.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	7,658[a]	5	,176
Razón de verosimilitudes	7,530	5	,184
Asociación lineal por lineal	2,420	1	,120
N de casos válidos	1797		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 17,14.

Tabla 6.35. Sigma asintótica para la prueba de Chi - Cuadrado.

Dado que la sigma asintótica = 0,176 > 0,05, se concluye que no hay diferencias significativas en las proporciones de los alumnos que completan primero o no en el primer año según su ubicación.

Para conocer si la ubicación de la escuela de procedencia de los alumnos es un factor relevante, se ha repetido el mismo análisis que se ha realizado para el lugar de residencia de los alumnos. Los resultados completos se pueden ver en el Anexo, pero a continuación se muestran las tablas más concluyentes:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	14,615[a]	4	,006
Razón de verosimilitudes	14,357	4	,006
Asociación lineal por lineal	10,822	1	,001
N de casos válidos	1437		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 19,33.

Tabla 6.36. Sigma asintótica para la prueba de Chi - Cuadrado para las variables Ubicacion_Escuela_Cod vs Grupos_Primero

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	8,609[a]	4	,072
Razón de verosimilitudes	8,477	4	,076
Asociación lineal por lineal	,040	1	,841
N de casos válidos	1795		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 24,53.

Tabla 6.37. Sigma asintótica para la prueba de Chi - Cuadrado para las variables Ubicacion_Escuela_Cod vs Completan_Primero_Al_Ritmo

En el caso de la variable "Grupos_Primero" se ha aplicado el análisis únicamente a aquellos estudiantes que han aprobado todos los créditos de primero, que son un total de 1439. Para la variable "Completan_Primero_Al_Ritmo" se ha realizado la prueba de Chi Cuadrado para todos los estudiantes de la base de datos salvo los excluidos por anomalías, quedando un total de 1797 alumnos.

Se puede observar en el primer caso una sigma asintótica de 0,006, por lo que la ubicación de la escuela es un parámetro que diferencia a aquellos alumnos que están por encima o por debajo de la media. En el segundo caso, la sigma asintótica es de 0,072. Este valor es ligeramente mayor al nivel de significación de 0,05, por lo que la ubicación de la escuela no es una variable a tener en cuenta a la hora de diferenciar los estudiantes que completan primero o no al ritmo marcado. No obstante, al ser un valor cercano al nivel de significación, se dejará que sea el propio modelo quien decida si la inclusión de dicha variable mejora o no la precisión de la predicción.

6.4. Análisis de primero en función de la nota de la selectividad

En este apartado se desea determinar si la nota de la selectividad es una buen factor para diferenciar entre los resultados obtenidos por los alumnos de primero. Con el objetivo de facilitar el análisis, la variable empleada será la de "Grupos_Sele". Tal y como se ha

explicado en el apartado descriptivo de la base de datos, esta variable codifica la nota de la selectividad de la siguiente manera:

- Entre un 5,00 y 6,99 → Grupos_Sele = 1
- Entre un 7,00 y 8,99 → Grupos_Sele = 2
- Entre un 9,00 y 10,99 → Grupos_Sele = 3
- Entre un 11,00 y 12,99 → Grupos_Sele = 4
- Más de un 13 → Grupos_Sele = 5

En primer lugar, resulta interesante observar si la nota promedio de primero varía en función del grupo de la selectividad asignado a cada estudiante:

Grupo Sele	Promedio Primero	N
1	1,73	19
2	4,48	17
3	5,61	834
4	6,34	898
5	7,58	43

Tabla 6.38. Promedios de primero en función del grupo de selectividad.

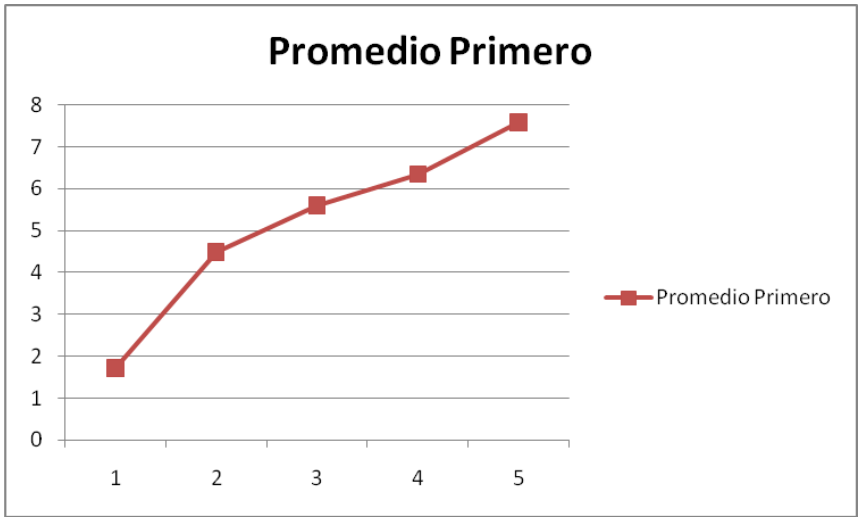

Figura 6.15. Promedios de primero en función del grupo de selectividad

Se observa una tendencia claramente ascendente. Es decir, cuanto mayor es la nota del grupo de la selectividad al que pertenece el alumno, mayor es la nota promedio obtenida en primero. No obstante, cabe mencionar que prácticamente el 96 % de los casos pertenecen a los grupos 3 y 4 (los que tienen una nota de la selectividad entre 9 y 12,99). El 4 % restante se divide entre el resto de grupos.

Dado que el promedio de primero varía considerablemente en función de los grupos, es evidente que el número de asignaturas repetidas también oscilará. En la siguiente tabla se muestra el número de asignaturas repetidas en promedio para cada uno de los grupos:

Grupo Sele	Promedio de Repetidas en Primero	N
1	6,74	19
2	4,82	17
3	4,84	834
4	1,99	898
5	0,07	43

Tabla 6.39. Promedio de repetidas por grupos de selectividad.

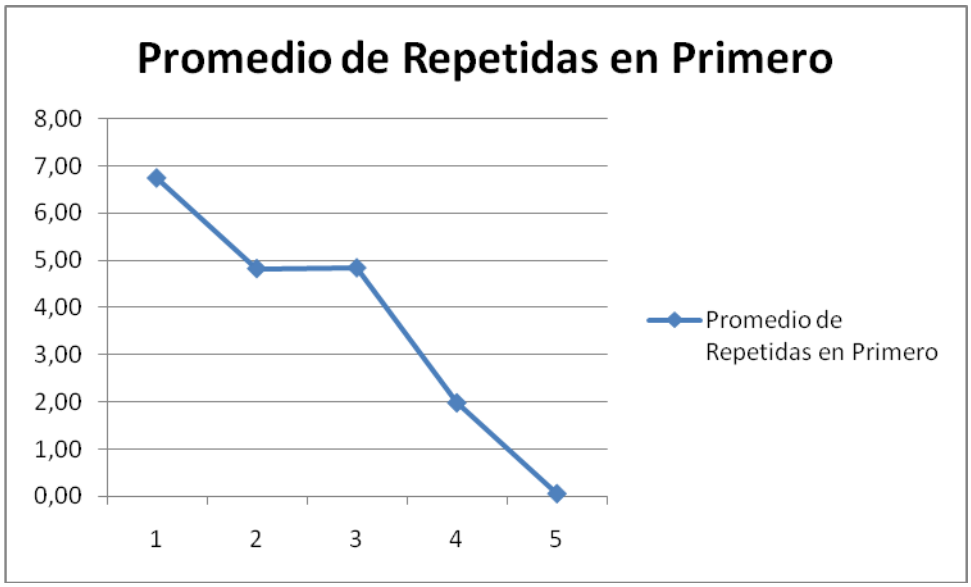

Figura 6.16. Promedio de repetidas por grupos de selectividad.

La tendencia en este caso es claramente descendente, dado que las personas con peor nota en la selectividad son las que repiten más asignaturas en primero (6,74 en promedio). Entre los grupos 2 y 3 apenas hay diferencia en el promedio de asignaturas repetidas. Sin embargo, entre los dos grupos con más datos (3 y 4), se aprecia una gran diferencia entre el promedio de suspendidas, ya que el grupo 2 suspende en promedio 4,84 asignaturas, y el grupo 4 en primedio suspende 1,99. Por último, existe un pequeño colectivo de 43 estudiantes que pertenecen al grupo cinco (más de un trece en la selectividad), que apenas suspenden en primero. El promedio total de asignaturas suspendidas por alumno en primero es de 3,33 asignaturas.

Resulta interesante comprobar como varía el porcentaje de los estudiantes que aprueban el primer curso en el primer año. Se muestra en la siguiente tabla cómo evoluciona este porcentaje en función de los grupos creados a partir de la nota de la selectividad.

Grupo Sele	Completan Primero Al ritmo
1	0,00%
2	0,00%
3	20,86%
4	54,68%
5	81,40%

Tabla 6.40. Porcentaje de alumnos que completan primero al ritmo en función de los grupos de la selectividad.

Los porcentajes para el grupo 1 y 2 son rotundos, ningún alumno perteneciente a alguno de estos dos grupos completa primero al ritmo marcado. En cuanto al grupo 5, sorprende que sólo el 81,40 % complete el primer curso en el primer año cuando el promedio de suspendidas es de 0,07. Se ha analizado este grupo en profundidad, y se ha concluido que la causa de que este porcentaje sea tan bajo es debido a estudiantes que se dejan alguna asignatura de primero para el segundo curso, pero durante el primer curso realizan asignaturas de segundo. Por este motivo no tienen el bloque de primero cerrado en el primer año, pero sin embargo, son alumnos que apenas repiten.

Se ha comprobado que existen diferencias en el promedio de la nota de primero en función de los grupos de la selectividad. Se desea observar ahora si estas diferencias se producen en todas las asignaturas de primero, y en caso de que las haya, mostrar las notas promedio de los grupos para cada una de las asignaturas. De este modo, es posible realizar una

estimación de las notas que sacará el alumno en cada una de las asignaturas teniendo en cuenta su nota de la selectividad.

Con el fin de aplicar las pruebas más adecuadas, es necesario en primer lugar saber si los datos siguen una distribución normal o no. Se ha aplicado la prueba de Kolmogorov - Smirnov a las notas de cada una de las asignaturas, segmentándolas por los grupos de la nota de la selectividad. La hipótesis nula de la prueba es que los datos siguen una distribución normal. El resultado del análisis es que en dos de los grupos (3 y 4), las notas de las asignaturas no siguen una distribución normal, por lo tanto se rechaza la hipótesis nula y es necesario recurrir a las pruebas no parámetricas. El resultado de la prueba Kolmogorov - Smirnov se encuentra en el Anexo.

Se desea determinar si existen diferencias en los promedios de cada una de las asignaturas de primero, y en caso afirmativo, saber si éstas son significativas. Para ello, se ha llevado a cabo la prueba no paramétrica de Kruskal Wallis para k grupos independientes bajo la hipótesis nula de que los datos provienen de la misma población. El parámetro más concluyente es la sigma asintótica y se muestra continuación:

Estadísticos de contraste[a,b]

	Algebra_ Lineal	Calculo_I	Mecanica_ Fundamental	Quimica_I	Fundamentos _Informatica
Chi-cuadrado	279,978	290,834	375,480	373,576	255,499
gl	4	4	4	4	4
Sig. asintót.	,000	,000	,000	,000	,000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Grupos_Sele

Tabla 6.41. Sigma asintótica para las asignaturas del primer cuatrimestre de primero

Estadísticos de contraste[a,b]

	Geometria	Calculo_II	Termodinamica_ Fundamental	Quimica_II	Expresion_ Grafica
Chi-cuadrado	175,633	92,018	186,321	171,767	99,275
gl	4	4	4	4	4
Sig. asintót.	,000	,000	,000	,000	,000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Grupos_Sele

Tabla 6.42. Sigma asintótica para las asignaturas del primer cuatrimestre de segundo

Dado que el valor de la sigma asintótica es inferior al nivel de significación de 0,05, se rechaza la hipótesis nula de que los datos provienen de la misma población y por lo tanto, se puede afirmar que existen diferencias significativas en los promedios de todas las asignaturas entre los distintos grupos de la selectividad. Verificado el hecho de que hay diferencias, se muestra a continuación cuáles son estos promedios por grupos para cada una de las asignaturas:

Grupos Sele	Algebra Lineal	Cálculo I	Mecánica Fundamental	Química I	Fundamentos Informática
1	1,49	1,58	1,67	1,65	1,46
2	3,06	3,76	3,89	3,64	3,23
3	5,11	5,27	5,08	5,44	5,26
4	6,12	6,22	6,07	6,54	6,50
5	7,49	7,87	7,63	8,18	8,16

Tabla 6.43. Promedios por grupos de las asignaturas del primer cuatrimestre de primero.

Grupos Sele	Geometría	Cálculo II	Termodinámica Fundamental	Química II	Expresión Gráfica
1	4,35	3,25	2,86	2,62	3,86
2	5,13	4,89	4,57	5,21	6,23
3	5,38	5,25	4,90	5,63	5,36
4	6,09	5,85	5,77	6,41	6,13
5	7,50	7,04	6,95	7,85	7,13

Tabla 6.44. Promedios por grupos de las asignaturas del segundo cuatrimestre de primero.

Llama especialmente la atención que en el grupo 1 y 2 los promedios de las asignaturas de primer cuatrimestre son mucho más bajos que los promedios de las asignaturas del segundo cuatrimestre. Con tal de encontrar una explicación a este hecho, se ha observado el número de datos con los que se han calculado los promedios para cada uno de los cuatrimestres. Se ha comprobado que existe un descenso en el número de datos empleados para calcular los promedios del segundo cuatrimestre. Mientras que en el primer cuatrimestre, tanto para el grupo 1 o como para el grupo 2, el número de observaciones para cada una de las asignaturas está entre 15 y 20, en el segundo cuatrimestre este número está comprendido

entre 3 y 10. Por lo tanto, la conclusión es clara, aquellos alumnos de primer y segundo grupo que obtienen unos resultados extremadamente malos, abandonan la titulación antes de llegar al segundo cuatrimestre.

Se observa que para todas las asignaturas salvo Expresión Gráfica, a grupos con mayor nota de selectividad, mayores son los promedios obtenidos en las asignaturas. El caso de Expresión Gráfica es anómalo, ya que se observa un promedio excesivamente alto para el grupo 2. La explicación que se da a este hecho es que la muestra con la que está calculada este promedio es únicamente de cinco estudiantes, y cualquier resultado algo extremo puede desviar el promedio fácilmente.

6.5. Modelo predictivo para los alumnos en primero

En este apartado se desea construir tres modelos que permitan prever el comportamiento del estudiante en primero en función de distintos parámetros. Las variables dependientes a predecir que definen el comportamiento del alumno en primero son:

- **Completan_Primero_Al_Ritmo** $\rightarrow$ Si aprueban todos los créditos de primero en el primer año toma el valor 1, sino toma el valor 0.
- **Repiten_Primero_S_N** $\rightarrow$ Si el alumno repite alguna asignatura en primero toma el valor 1, si no repite ninguna toma el valor 0.
- **Grupos_Primero** $\rightarrow$Si el alumno está por encima de la media de primero toma el valor 1, si está por debajo toma el valor 0.

La técnica empleada para la creación de los modelos y la posterior clasificación de los casos es la regresión logística hacia adelante por el método de Wald. Se ha descartado el uso de otros métodos como la regresión lineal debido a que alguna de las variables independientes son cualitativas. Asimismo, se ha empleado el uso del análisis discriminante como técnica de clasificación, pero se ha optado por la regresión logística debido a que ofrece un mayor porcentaje de acierto. Los resultados del análisis discriminante se pueden observar en el anexo.

Para la creación de los modelos es necesario dividir los datos en dos subconjuntos: uno de entrenamiento para la creación del modelo (90 % de los datos) y otro de validación para verificar la precisión del mismo (10 % de los datos). Es especialmente importante para garantizar la validez del modelo que el subconjunto de la validación no participe en la creación de éste, ya que este hecho podría dar lugar a una precisión del modelo por encima de la real.

Para realizar la regresión logística, en primer lugar se deben seleccionar aquellas variables potencialmente predictoras. Una vez seleccionadas, existen distintos métodos para determinar cuáles participaran finalmente en el modelo: se puede seleccionar un método que incluya en el modelo todas las variables seleccionadas (método introducir); se puede partir de cero e ir añadiendo las variables por pasos una a una en función de su significancia (método hacia adelante); o se puede partir de que todas las variables participan en el modelo, e ir eliminando de una en una y por pasos aquellas que no aportan precisión al modelo (método hacia atrás). El método utilizado en este proyecto es el método hacia adelante de Wald, que se basa en este estadístico (Wald) para ir introduciendo las variables en el modelo.

Para prever el comportamiento de primero se disponen de las siguientes variables independientes:

- Sexo (SexoCod)
- Lugar de residencia (UbicacionCodif)
- Ubicación de la escuela donde ha estudiado el alumno (Ubicacion_Escuela_Cod)
- Nota de la selectividad (Nota_Sele)

De estas cuatro variables únicamente se tendrán en cuenta tres de ellas, ya que en capítulos anteriores se ha visto que el sexo no es una buena variable predictora, ya que no existen diferencias significativas entre hombres y mujeres. De este modo, las posibles variables introducidas en los modelo serán tres: UbicacionCodif, Ubicacion_Escuela_Cod, y Nota_Sele.

6.5.1. Predicción de los estudiantes que aprobarán primero en el primer año

Con este modelo se desea predecir qué alumnos aprobarán todos los créditos del primer curso en un año y cuáles no. La estructura de datos es la siguiente:

- Datos empleados para la creación del modelo = 1617 (90,0 %)
- Datos empleados para la validación del modelo = 178 (9,9 %)
- Datos perdidos = 2 (0,1 %)
- Datos totales = 1797

Codificación de la variable dependiente:

- Completan_Primero_Al_Ritmo = 0 $\rightarrow$ No completan primero en el primer año.
- Completan_Primero_Al_Ritmo = 1 $\rightarrow$ Completan primero en el primer año.

Variables preseleccionadas para la creación del modelo:

- Nota_Sele
- UbicacionCodif
- Ubicación_Escuela_Cod

Los resultados mostrados por el programa SPSS se dividen en distintos bloques. El primero de ellos se denomina como Bloque 0, y en él se muestran los resultados obtenidos aplicando el modelo predictivo más básico, que consiste en clasificar todos los casos en el grupo de mayor frecuencia. El objetivo de este bloque es demostrar cómo se consigue una mayor precisión del modelo realizando una correcta calibración de los coeficientes β_i.

A continuación se muestra una primera clasificación de los datos:

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Completan_Primero_Al_Ritmo		Porcentaje correcto	Completan_Primero_Al_Ritmo		Porcentaje correcto
	Observado		0	1		0	1	
Paso 0	Completan_Primero_Al_Ritmo	0	983	0	100,0	113	0	100,0
		1	634	0	,0	65	0	,0
	Porcentaje global				60,8			63,5

Tabla 6.45. Clasificación realizada en el bloque 0.

Se observa que un porcentaje de los estudiantes ligeramente superior al 60 % no completan primero al ritmo, por lo que clasificando todos los estudiantes en el grupo 0, se obtienen unos porcentajes de acierto del 60,8 % en el subconjunto de entrenamiento, y del 63,5 % en el subconjunto de validación.

En la siguiente tabla se puede observar como ninguna de las variables preseleccionadas forma parte del modelo, ya que éste está constituido únicamente por una constante. Es importante fijarse en el parámetro sigma, que indica la significancia de cada una de las variables a la hora de diferenciar entre los grupos de la variable dependiente. Cuanto más cercana a cero es la sigma, más relevante es esa variable.

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	UbicacionCodif	2,007	1	,157
		Ubicacion_Escuela_Cod	,000	1	,984
		Nota_Sele	255,189	1	,000
	Estadísticos globales		256,865	3	,000

Tabla 6.46. Significancia de las variables preseleccionadas.

Se puede apreciar que únicamente la variable "Nota_Sele" tiene una sigma menor al nivel de significación marcado de 0,05, por lo que el modelo será univariante.

El procedimiento que se sigue para realizar la regresión logística mediante el método hacia adelante de Wald se muestra en el siguiente bloque de resultados ofrecido por el programa SPSS, denominado Bloque 1. En él se muestran los distintos resultados obtenidos en cada uno de los pasos realizados en la regresión. En cada uno de estos pasos se añade una de las variables explicativas preseleccionadas para el modelo, y se comprueba si ésta aumenta la precisión de éste. En el caso del modelo para ver si los estudiantes completan primero al ritmo o no, el modelo consta de un único paso, ya que sólo una de las variables preseleccionadas es significativa, y ésta es la única que se introduce.

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	313,195	1	,000
	Bloque	313,195	1	,000
	Modelo	313,195	1	,000

Tabla 6.47. Sigma del modelo

Se observa un sigma del modelo de 0,000, hecho que indica que el modelo constituido por una única variable (Nota_Sele) es significativo. A continuación se muestra una tabla con el valor de la R Cuadrado de Nagelkerke, que indica qué porcentaje de la varianza es explicado al introducir las variables en el modelo. Como se muestra en la siguiente tabla, en este caso la nota de la selectividad explica un 23,9 % de la varianza total de los datos.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1852,522(a)	,176	,239

Tabla 6.48. Valor de la R Cuadrado de Nagelkerke

Resulta especialmente interesante observar si el modelo creado se ajusta a los datos observados. Para ello se aplica la prueba de Hosmer y Lemeshow, que es el test de bondad de ajuste más empleado en las regresiones logísticas. Con el test de Hosmer y Lemeshow, se dividen los datos en intervalos, y para cada uno de estos se mide la distancia entre lo observado y lo esperado bajo el modelo. El estadístico empleado para este test es el de Chi - Cuadrado y el parámetro más relevante a observar es la sigma asintótica. En la siguiente tabla se puede ver su valor:

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	10,869	8	,209

Tabla 6.49. Sigma de la prueba de Hosmer y Lemeshow

Dado que el valor de la sigma es mayor que el nivel de significación de 0,05, se puede afirmar que el modelo se ajusta bien a los datos observados.

A continuación se muestran las variables que participan en el modelo, que este caso sólo es una:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1(a)	Nota_Sele	1,077	,070	236,164	1	,000	2,935
	Constante	-12,505	,792	249,579	1	,000	,000

a Variable(s) introducida(s) en el paso 1: Nota_Sele.

Tabla 6.50. Variables que constituyen el modelo

Se observa que el valor de la sigma es de 0,000, hecho que comporta que la variable "Nota_Sele" es significativa a la hora de discernir entre grupos. Otro valor importante que se puede apreciar en la tabla anterior es del parámetro B. A medida que este parámetro es más grande, indica que cuanto mayor es el valor de esa variable independiente, más contribuye a que la variable dependiente adopte el valor 1 (complete primero al ritmo estipulado en este caso). Por el contrario un valor del parámetro B negativo indica que cuánto más grande es el valor de esa variable independiente más contribuye a que la variable dependiente adopte el valor 0 (no complete primero en un año en este caso). Dado que el valor de B para la variable "Nota_Sele" es positivo en este caso, se puede afirmar que cuanto mayor es la nota de la selectividad, más probable es que el alumno complete el primer curso en el primer año.

Finalmente el programa SPSS muestra una tabla de clasificación de los casos empleando el modelo creado. Se puede ver a continuación:

Tabla de clasificación(c)

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Completan_Primero_Al_Ritmo		Porcentaje correcto	Completan_Primero_Al_Ritmo		Porcentaje correcto
	Observado		0	1		1	0	
Paso 1	Completan_Primero_Al_Ritmo	0	830	153	84,4	97	16	85,8
		1	309	325	51,3	36	29	44,6
	Porcentaje global				71,4			70,8

Tabla 6.51. Clasificación realizada por el modelo generado

El porcentaje de acierto para el subconjunto de datos con los que se ha creado el modelo es de 71,4 %, y para el subconjunto de validación es de un 70,8 %. Es decir, para un estudiante nuevo, conociendo su nota de la selectividad se podrá predecir si completará o no el primer curso durante el primer año con un 70,8 % de seguridad.

La probabilidad de que el alumno complete primero al ritmo marcado es la siguiente:

$$P(Comple\tan\_\ \Pr imero\_Al\_Ritmo=1)=\frac{1}{1+e^{(12{,}505-1{,}077\cdot Nota\_Sele)}}$$

El punto de corte para la clasificación en grupos es 0,5. Si $P(Comple\tan\_\ \Pr imero\_Al\_Ritmo=1)$ es mayor o igual que 0,5, se clasifica al alumno en el grupo 1, que son aquellos estudiantes que completan primero al ritmo marcado.

6.5.2. Predicción de los alumnos que estarán por encima de la media de primero

El modelo creado en este apartado servirá para predecir qué alumnos estarán por encima de la media de primero y qué alumnos estarán por debajo. Para calcular la media, se ha seleccionado únicamente aquellos alumnos que han completado primero, que son un total de 1439. La nota media en primero de dichos alumnos es de 6,35. La estructura de datos para la creación del modelo es la siguiente:

- Datos empleados para la creación del modelo = 1293 (89,9 %)
- Datos empleados para la validación del modelo = 144 (10,0 %)
- Datos perdidos = 2 (0,1 %)
- Datos totales = 1439

Codificación de la variable dependiente:

- Grupos_Primero = 0 $\rightarrow$ Tienen una nota media inferior a la media de primero (6,35).
- Grupos_Primero = 1 $\rightarrow$ Tienen una nota media igual o superior a la media de primero (6,35).

Variables preseleccionadas para la creación del modelo:

- Nota_Sele
- UbicacionCodif
- Ubicación_Escuela_Cod

Clasificando todos los casos en el grupo de mayor frecuencia (que es el de los estudiantes por debajo de la media de primero), se obtienen unos porcentajes de acierto alrededor del 60 %:

Tabla de clasificación(c,d)

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Grupos_Primero		Porcentaje correcto	Grupos_Primero		Porcentaje correcto
	Observado		0	1		0	1	
Paso 0	Grupos_Primero	0	784	0	100,0	86	0	100,0
		1	509	0	,0	58	0	,0
	Porcentaje global				60,6			59,7

Tabla 6.52. Clasificación del bloque 0

Siguiendo la misma pauta que en el modelo anterior, se observa ahora la significancia de las variables preseleccionadas:

			Puntuación	gl	Sig.
Paso 0	Variables	UbicacionCodif	14,640	1	,000
		Ubicacion_Escuela_Cod	8,315	1	,004
		Nota_Sele	247,887	1	,000
	Estadísticos globales		259,706	3	,000

Tabla 6.53. Significación de las variables preseleccionadas

A diferencia del modelo anterior, se observa en este caso como las tres variables preseleccionadas tienen una sigma menor a 0,05, hecho que significa que las tres son significativas.

El modelo para predecir si el alumno estará por encima o por debajo de la media de primero consta de dos variables que se introducen en dos pasos:

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1(a)	Nota_Sele	1,231	,085	211,137	1	,000	3,425
	Constante	-14,372	,966	221,467	1	,000	,000
Paso 2(b)	UbicacionCodif	,188	,052	13,000	1	,000	1,207
	Nota_Sele	1,234	,085	210,089	1	,000	3,435
	Constante	-14,696	,978	225,658	1	,000	,000

a Variable(s) introducida(s) en el paso 1: Nota_Sele.
b Variable(s) introducida(s) en el paso 2: UbicacionCodif.

Tabla 6.54. Variables que constituyen el modelo.

En el primero se introduce la variable "Nota_Sele", de la cual se puede decir que cuanto mayor es más probable es que el alumno pertenezca al grupo de los que están por encima de la media. En el segundo paso se introduce la variable "UbicacionCodif", que hace referencia al lugar de residencia del estudiante. Debido a que esta variable es categórica, el valor de B es cercano a 0, ya que aunque esta variable ayuda a distinguir entre grupos, no tiene una tendencia.

A continuación se muestra el porcentaje de varianza explicado en cada uno de los pasos, que corresponde al valor de la R Cuadrado de Nagelkerke:

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1456,919(a)	,193	,261
2	1443,911(a)	,201	,272

Tabla 6.55. Valor de la R Cuadrado de Nagelkerke

Con la introducción de la primera variable (Nota_Sele), se explica un 26,1 % de la varianza total de los datos, y con la introducción de la segunda variable (UbicacionCodif) se aumenta hasta un 27,2 %.

Aplicamos la prueba de bondad de ajuste de Hosmer y Lemeshow para observar si el modelo creado se adapta a los datos reales:

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	14,858	8	,062
2	18,561	8	,017

Tabla 6.56. Sigma de la prueba de Hosmer y Lemeshow

En este punto se ha detectado una anomalía en los resultados. Se observa una sigma ligeramente superior a 0,05 en el primero caso, por lo que se puede afirmar que el modelo se amolda de forma correcta a la realidad. Sin embargo, al introducir la segunda variable, la sigma toma el valor 0,017, hecho que indica que el modelo no se adapta a los datos reales. No obstante, se ha repetido la regresión logística introduciendo únicamente la variable "Nota_Sele" en el modelo para observar los resultados. Se ha podido comprobar que la

precisión obtenida con el modelo univariante es inferior a la obtenida con el modelo de dos variables cuya sigma es inferior a 0,05 (un 72,2 % de acierto frente a un 70,8 %). Por este motivo, el modelo empleado para la clasificación de casos es el de dos variables. Los resultados obtenidos tanto en el subconjunto de entrenamiento como en el subconjunto de validación son los siguientes:

Tabla de clasificación(c)

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Grupos_Primero		Porcentaje correcto	Grupos_Primero		Porcentaje correcto
	Observado		0	1		0	1	
Paso 1	Grupos_Primero	0	656	128	83,7	73	13	84,9
		1	230	279	54,8	29	29	50,0
	Porcentaje global				72,3			70,8
Paso 2	Grupos_Primero	0	665	119	84,8	75	11	87,2
		1	229	280	55,0	29	29	50,0
	Porcentaje global				73,1			72,2

Tabla 6.57. Tabla de clasificación empleando el modelo generado.

Como es lógico, el porcentaje de acierto en el subconjunto de entrenamiento es ligeramente al del subconjunto de validación, pero es este último el que se acerca más a la realidad, ya estos casos no han intervenido en la creación del modelo.

La expresión dada por el modelo para determinar la probabilidad de que el alumno esté por encima de la media de primero es la siguiente:

$$P(Grupos\_Primero=1)=\frac{1}{1+e^{(14{,}696-1{,}234\cdot Nota\_Sele-0{,}188\cdot UbicacionCodif)}}$$

Si $P(Grupos\_Primero=1)$ es mayor o igual que 0,5, se clasifica al alumno en el grupo 1, que son aquellos alumnos por encima de la media de primero.

6.5.3. Predicción de los alumnos que repetirán alguna asignatura en primero

En este apartado se desea predecir qué alumnos repetirán alguna asignatura en primero y cuáles no. Para la creación del modelo en este caso, se han tenido en cuenta todos los estudiantes de la base de datos (hayan completado primero o no) salvo los que presentan alguna anomalía. El resumen de los datos utilizados es el siguiente:

- Datos empleados para la creación del modelo = 1630 (90,0 %)
- Datos empleados para la validación del modelo = 179 (9,9 %)
- Datos perdidos = 2 (0,1 %)
- Datos totales = 1811

Codificación de la variable dependiente:

- Repiten_Primero_S_N = 0 → No repiten ninguna asignatura.
- Repiten_Primero_S_N = 1 → Repiten una o más asignaturas.

Variables preseleccionadas para la creación del modelo:

- Nota_Sele
- UbicacionCodif
- Ubicación_Escuela_Cod

En este caso, el grupo de mayor frecuencia es el de los estudiantes que repiten alguna asignatura en primero (Repiten_Primero_S_N = 1), que corresponde aproximadamente a un 65 % de los estudiantes.

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Repiten_Primero_S_N		Porcentaje correcto	Repiten_Primero_S_N		Porcentaje correcto
	Observado		0	1		0	1	
Paso 0	Repiten_Primero_S_N	0	0	572	,0	0	61	,0
		1	0	1058	100,0	0	118	100,0
	Porcentaje global				64,9			65,9

Tabla 6.58. Clasificación realizada en el bloque 0.

La significancia de la tres variables preseleccionadas es la siguiente:

			Puntuación	gl	Sig.
Paso 0	Variables	UbicacionCodif	4,426	1	,035
		Ubicacion_Escuela_Cod	1,593	1	,207
		Nota_Sele	311,084	1	,000
	Estadísticos globales		318,830	3	,000

Tabla 6.59. Significancia de las variables preseleccionadas

Salvo la ubicación de la escuela donde ha estudiado el alumno (Ubicacion_Escuela_Cod), las otras dos variables (Nota_Sele y UbicacionCodif) son relevantes a la hora de determinar si un alumno repetirá en primero o no.

Las variables introducidas en el modelo por el método hacia adelante de Wald son las siguientes:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1(a)	Nota_Sele	-1,279	,076	283,242	1	,000	,278
	Constante	15,009	,864	301,591	1	,000	3298414,017

a Variable(s) introducida(s) en el paso 1: Nota_Sele.

Tabla 6.60. Variables introducidas en el modelo

Como se puede observar, en este caso vuelve a haber un sólo paso, por lo que estamos frente a otro modelo univariante. La variable introducida es la nota de la selectividad, y tal y como indica un valor negativo de B, cuanto más alta es la nota en la selectividad más probable es que el alumno pertenezca al grupo 0 (alumnos que no repiten ninguna asignatura).

El porcentaje de varianza explicado lo marca el valor de la R Cuadrado de Nagelkerke, y es de 29,7 %:

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1716,396(a)	,216	,297

Tabla 6.61. Valor de la R Cuadrado de Nagelkerke

Siguiendo los mismos pasos que en el apartado anterior, se muestra ahora los resultados obtenidos al aplicar la prueba de bondad de ajuste de Hosmer y Lemeshow:

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	6,905	8	,547

Tabla 6.62. Sigma de la prueba de Hosmer y Lemeshow

Se observa un valor de 0,547, que es superior al nivel de significación de 0,05, por lo que el modelo se ajusta bien a los datos observados. La tabla de clasificación perteneciente al modelo es la siguiente:

Tabla de clasificación(c)

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Repiten_Primero_S_N		Porcentaje correcto	Repiten_Primero_S_N		Porcentaje correcto
	Observado		0	1		0	1	
Paso 1	Repiten_Primero_S_N	0	296	276	51,7	26	35	42,6
		1	135	923	87,2	15	103	87,3
	Porcentaje global				74,8			72,1

Tabla 6.63. Tabla de clasificación utilizando el modelo generado

Por lo tanto, con el modelo construido se podrá pronosticar qué alumnos repetirán en primero o no con un 72,1 de acierto.

La ecuación del modelo que determina la probabilidad de que el alumno repita en primero es la siguiente:

$$P(\text{Re}\ piten\_\text{Pr}imero\_S\_N=1)=\frac{1}{1+e^{(15{,}009-1{,}279\cdot Nota\_Sele)}}$$

Si $P(\text{Re}\ piten\_\text{Pr}imero\_S\_N=1)$ es mayor o igual que 0,5, se clasifica al alumno se pronostica que el alumno repetirá alguna asignatura en primero.

6.6. Modelo predictivo para los alumnos en segundo

Tal y como se ha realizado con el primer curso, en este apartado se desean construir tres modelos predictivos para determinar el comportamiento de los alumnos en segundo. Para ello, se volverá a aplicar una regresión logística, que es la técnica que ofrece mejores resultados, con el fin de clasificar a los alumnos entre los distintos grupos de segundo existentes: si están por encima o por debajo de la media en segundo (Grupos_Segundo), si repiten o no en segundo (Repiten_Segundo_S_N), o si completan segundo al ritmo estipulado o no (Completan_Segundo_Al_Ritmo). El método de introducción de variables en el modelo será de nuevo el método hacia adelante de Wald, para el cual se debe preseleccionar previamente un abanico de variables potencialmente predictoras. Una vez seleccionadas, la metodología es la siguiente: se parte de la base de que el modelo es una constante, y se van añadiendo las variables significativas en pasos sucesivos para poder clasificar a los estudiantes en los distintos grupos. Para determinar si una variable es significativa o no, se utiliza el estadístico de probabilidad de Wald.

6.6.1. Predicción de los estudiantes que aprobarán segundo en el segundo año de carrera

El primer modelo de segundo a confeccionar será para determinar qué alumnos completan segundo en su segundo año de grado y cuáles no (Completan_Segundo_Al_Ritmo). El primer paso a realizar consiste en filtrar los datos con el objetivo de separar aquellos que puedan llevar a conclusiones erróneas. En este caso, se deben eliminar de la muestra todos aquellos alumnos que han ingresado en la universidad en 2013, ya que ninguno de éstos ha completado segundo yendo al ritmo estipulado de curso por año (se poseen datos hasta el primer cuatrimestre de 2014). Aplicado este filtro, es necesario dividir los datos en dos subconjuntos: uno de entrenamiento para crear el modelo, y otro de validación para verificar la precisión del modelo. Es importante remarcar que los datos de éste último subconjunto no participan en la creación del modelo. La estructura de datos es la siguiente:

- Datos empleados para la creación del modelo = 1228 (90,4 %)
- Datos empleados para la validación del modelo = 131 (9,6 %)

- Datos totales = 1359

Codificación de la variable dependiente:

- Completan_Segundo_Al_Ritmo = 0 → No completan segundo en el segundo año de carrera.
- Completan_Segundo_Al_Ritmo = 1 → completan segundo en el segundo año de carrera.

Variables preseleccionadas para la creación del modelo:

- Nota_Sele
- Promedio_Primero
- Grupos_Primero
- Repetidas_Primero
- Alpha_Primero
- Completan_Primero_Al_Ritmo

El grupo de mayor frecuencia es el Completan_Segundo_Al_Ritmo = 0, que representa un 73,3 % de los casos. Por lo tanto, elaborando el modelo con una constante que clasifique a todos los estudiantes en dicho grupo, se conseguirá un 73,3 % de acierto.

			Pronosticado					
			Casos seleccionados(a)			Casos no seleccionados(b)		
			Completan_Segundo_Al_Ritmo		Porcentaje correcto	Completan_Segundo_Al_Ritmo		Porcentaje correcto
	Observado		0	1		0	1	
Paso 0	Completan_Segundo_Al_Ritmo	0	907	0	100,0	96	0	100,0
		1	321	0	,0	35	0	,0
	Porcentaje global				73,9			73,3

Tabla 6.64. Primera clasificación del modelo del Bloque 0.

A continuación, se muestra la significación de las variables preseleccionadas para observar si distinguen bien entre grupos o no. El parámetro más importante es la sigma, que si adopta un valor inferior al nivel de significación de 0,05 indica que esa variable sí que diferencia bien entre grupos.

			Puntuación	gl	Sig.
Paso 0	Variables	Nota_Sele	150,582	1	,000
		Promedio_Primero	259,571	1	,000
		Grupos_Primero	,413	1	,520
		Repetidas_Primero	298,796	1	,000
		Alpha_Primero	317,427	1	,000
		Completan_Primero_Al_Ritmo	432,494	1	,000
	Estadísticos globales		548,391	6	,000

Tabla 6.65. Significancia de las variables preseleccionadas.

Se observa como a priori, todas las variables son significantes para construir el modelo salvo la variable "Grupos_Primero", que presenta una sigma asintótica mayor al nivel de significación marcado de 0,05. Dicha variable codifica a los alumnos con un promedio superior o inferior a la media en primero y a priori, no es relevante para confeccionar el modelo.

El modelo generado para predecir si un completará segundo al ritmo o no consta de 6 pasos, en los que se han ido introduciendo y quitando variables hasta obtener un modelo final con 4 variables, que son: Promedio_Primero, Grupos_Primero, Repetidas_Primero, y Completan_Primero_Al_Ritmo. En el Anexo puede observarse cómo en qué orden se han ido introduciendo las variables en el modelo paso a paso. Se puede comprobar cómo finalmente la variable "Grupos_Primero", cuya sigma era superior a 0,05 (variable no significativa), entra en el modelo en el último paso. La razón se puede deber a que puede ser una variable mala para predecir desde un inicio, pero con un modelo ya hecho (en el paso 6 el modelo ya tiene 3 variables), sí que puede ayudar a clasificar algunos casos correctamente, mejorando así la precisión del modelo. La sigma del modelo en cada uno de los pasos se muestra a continuación:

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	488,641	1	,000
	Bloque	488,641	1	,000
	Modelo	488,641	1	,000
Paso 2	Paso	219,190	1	,000
	Bloque	707,830	2	,000
	Modelo	707,830	2	,000
Paso 3	Paso	71,335	1	,000
	Bloque	779,166	3	,000
	Modelo	779,166	3	,000
Paso 4	Paso	7,750	1	,005
	Bloque	786,916	4	,000
	Modelo	786,916	4	,000
Paso 5[a]	Paso	-,140	1	,709
	Bloque	786,776	3	,000
	Modelo	786,776	3	,000
Paso 6	Paso	7,017	1	,008
	Bloque	793,793	4	,000
	Modelo	793,793	4	,000

Tabla 6.66. Significancia de los modelos en cada uno de los pasos

Se puede apreciar una sigma del modelo de 0,000 en todos los pasos de la regresión. Al ser inferior al nivel de significación de 0,05, se concluye que el modelo obtenido al ir introduciendo las variables en cada uno de los pasos es significante. A continuación se muestra el valor de la R cuadrado de Nagelkerke, que corresponde al porcentaje de la varianza explicado en cada uno de los pasos. Por lo tanto, explica en qué medida mejora el modelo con las nuevas variables.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	922,373[a]	,328	,481
2	703,183[b]	,438	,641
3	631,848[c]	,470	,688
4	624,097[d]	,473	,693
5	624,237[e]	,473	,693
6	617,220[e]	,476	,697

Tabla 6.67. Valores de la R cuadrado de Nagelkerke al introducir las variables.

Tal y como se observa en la tabla, sólo con la primera variable introducida en el modelo (Completan_Primero_Al_Ritmo) se explica ya el 48,1 % de la varianza de los casos con los que se genera el modelo. Al introducir la segunda variable (Promedio_Primero), ya se aumenta hasta un 64,1 %. Finalmente, con todas las variables del modelo (Completan_Primero_Al_Ritmo, Promedio_Primero, Grupos_Primero y Repetidas_Primero).

Con el modelo ya construido, se aplica la prueba de bondad de ajuste de Hosmer y Lemeshow para observar si se ajusta de manera significativa a los datos reales:

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	0	.
2	36,825	8	,000
3	24,817	8	,002
4	10,579	8	,227
5	8,831	8	,357
6	4,613	8	,798

Tabla 6.68. Valores de Chi - Cuadrado y de la sigma asintótica para la prueba de Hosmer y Lemeshow

Se observa como a medida que se van introduciendo variables en los pasos, el valor de la sigma se va incrementando. El modelo obtenido en el cuarto paso ya se ajusta de manera significativa a la realidad, por lo que el obtenido en pasos sucesivos también.

Las variables que intervienen en la ecuación junto a sus coeficientes B se pueden observar a continuación:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 6[a]	Promedio_Primero	1,755	,223	61,863	1	,000	5,782
	Grupos_Primero	-,433	,169	6,575	1	,010	,649
	Repetidas_Primero	-1,172	,180	42,602	1	,000	,310
	Completan_Primero_Al_Ritmo	1,192	,281	18,045	1	,000	3,294
	Constante	-11,839	1,476	64,359	1	,000	,000

Tabla 6.69. Variables del modelo con sus coeficientes B.

Se observa que para todas ellas el valor de la sigma es de inferior a 0,05, hecho que comporta que todas ellas son significativas a la hora de discernir entre grupos. En cuanto al valor del parámetro B, se puede deducir que, promedios elevados en primero y alumnos que han completado el primer curso en el primer año tienen más probabilidad de completar el segundo curso en el segundo año. Sin embargo, se observa una contradicción con lo que sucede con la variable "Grupos_Primero", ya que indica que los alumnos por debajo de la media en primero completarán segundo al ritmo marcado más fácilmente que los que están por encima. Esta información es totalmente contraria a lo que marca el modelo con la variable "Promedio_Primero", pero no se dará especial importancia a este hecho debido a la diferencia de escalas en los valores de B (para la variable "Grupos_Primero" se tiene un valor cercano a 0, por lo que el peso de dicha variable es muy bajo). También se puede observar un valor negativo de B para la variable "Repetidas_Primero", por lo que a más asignaturas repetidas en primero, menos probabilidad de completar segundo en el segundo año de carrera.

La clasificación final obtenida aplicando el modelo al subconjunto de entrenamiento y al subconjunto de validación es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Completan_Segundo_Al_Ritmo		Porcentaje correcto	Completan_Segundo_Al_Ritmo		Porcentaje correcto
	Observado		0	1		0	1	
Paso 6	Completan_Segundo_Al_Ritmo	0	811	96	89,4	82	14	85,4
		1	60	261	81,3	6	29	82,9
	Porcentaje global				87,3			84,7

Tabla 6.70. Clasificación realizada por el modelo.

Se puede observar cómo el modelo mejora el porcentaje de acierto de inicio, consiguiendo hasta un 87,3 % en el subconjunto de entrenamiento, y un 84,7 % en el subconjunto de validación. Por lo tanto, la probabilidad clasificar en el grupo correcto un nuevo estudiante es del 84,7 % (porcentaje del subconjunto de validación).

La probabilidad de que el alumno complete segundo en el segundo año de carrera es la siguiente:

$$P(Completan\_Segundo\_Al\_Ritmo=1)=\frac{1}{1+e^{\begin{array}{l}(11{,}839-1{,}192\cdot Completan\_Primero\_Al\_Ritmo+\\+1{,}172\cdot Repetidas\_Primero+0{,}433\cdot Grupos\_Primero+\\(-1{,}755)\cdot Promedio\_Primero)\end{array}}}$$

El punto de corte para la clasificación en grupos es 0,5. Si $P(Comple\tan\_\ Segundo\_Al\_Ritmo = 1)$ es mayor o igual que 0,5, se clasifica al alumno en el grupo 1, que son aquellos estudiantes que completan segundo al ritmo marcado.

6.6.2. Predicción de los alumnos que estarán por encima de la media de segundo

En este apartado se mostrará el modelo creado para predecir qué estudiantes de segundo estarán por encima de la media en dicho curso, y cuáles no. Para evitar datos que distorsionen el modelo, es especialmente importante realizar un correcto filtrado de los datos con los que se creará el modelo. En este caso, se ha construido y validado el modelo únicamente con aquellos estudiantes que han completado segundo, que son 609. Asimismo, la media de segundo para crear los grupos se ha calculado también con los estudiantes que han completado dicho curso, obteniendo una media de 6,57. El reparto de datos queda de la siguiente manera:

- Datos empleados para la creación del modelo = 549 (90,1 %)
- Datos empleados para la validación del modelo = 60 (9,9 %)
- Datos totales = 609

Codificación de la variable dependiente:

- Grupos_Segundo = 0 $\rightarrow$ Alumnos con una nota promedio en segundo inferior a 6,57.
- Grupos_Segundo = 1 $\rightarrow$ Alumnos con una nota promedio en segundo igual o superior a 6,57.

Variables preseleccionadas para la creación del modelo:

- Nota_Sele
- Promedio_Primero
- Grupos_Primero
- Repetidas_Primero

- Alpha_Primero
- Completan_Primero_Al_Ritmo

La clasificación realizada por el programa en el bloque 0 es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Grupos_Segundo		Porcentaje correcto	Grupos_Segundo		Porcentaje correcto
	Observado		0	1		0	1	
Paso 0	Grupos_ Segundo	0	319	0	100,0	30	0	100,0
		1	230	0	,0	30	0	,0
	Porcentaje global				58,1			50,0

Tabla 6.71. Clasificación realizada en el bloque 0

Se observa cómo para el subconjunto de validación, un 50 % de los casos están por encima de la media de segundo y un 50 % están por debajo. Sin embargo, para el subconjunto de entrenamiento predominan los alumnos por debajo de la media, que son un 58,1 %.

El modelo se ha creado en dos pasos en este caso, estando formado únicamente por dos variables de las seis preseleccionadas:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Promedio_Primero	2,433	,211	132,590	1	,000	11,392
	Constante	-16,445	1,401	137,706	1	,000	,000
Paso 2[b]	Nota_Sele	,675	,146	21,421	1	,000	1,965
	Promedio_Primero	2,091	,218	92,401	1	,000	8,093
	Constante	-21,885	1,952	125,684	1	,000	,000

Tabla 6.72. Variables del modelo con sus coeficientes B

En el primer paso, se introduce la variable "Promedio_Primero", de la cual observando su parámetro B, se puede afirmar que cuanto mayor es el promedio en primero, más fácil es que el alumno pertenezca al grupo 1 (alumnos por encima de la media en segundo). Lo mismo pero en menor escala sucede con la variable "Nota_Sele": a mayor nota de selectividad, más probabilidad de tener una nota igual o superior a 6,57 en segundo.

A continuación se muestran los valores de la R Cuadrado de Nagelkerke y de la prueba de bondad de ajuste de Hosmer y Lemeshow:

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	509,848[a]	,350	,471
2	487,085[a]	,377	,507

Tabla 6.73. Valor de la R Cuadrado de Nagelkerke

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	4,429	8	,816
2	3,988	8	,858

Tabla 6.74. Sigma de la prueba de Hosmer y Lemeshow

Se puede observar como el valor de la sigma de la prueba de Hosmer y Lemeshow es superior a 0,05, por lo que el modelo creado se ajusta de manera significativa a los datos reales, explicando un porcentaje de varianza del 50,7 %, que es el valor adoptado por la R Cuadrado de Nagelkerke.

La tabla de clasificación final aplicando el modelo es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Grupos_Segundo		Porcentaje correcto	Grupos_Segundo		Porcentaje correcto
	Observado		0	1		0	1	
Paso 2	Grupos_Segundo	0	273	46	85,6	27	3	90,0
		1	67	163	70,9	15	15	50,0
	Porcentaje global				79,4			70,0

Tabla 6.75. Tabla de clasificación empleando el modelo generado

Prácticamente un 80 % de los casos se clasifican con éxito en el subconjunto de entrenamiento. Este valor es inferior en el subconjunto de entrenamiento, que es del 70 %. Se observa sólo un 50 % de acierto en el modelo en aquellos alumnos que pertenecen al grupo 1 (alumnos por encima de la media), y que el modelo los clasifica en el grupo 0 (alumnos por debajo de la media). Por el contrario, los alumnos pertenecientes al grupo 0 son detectados con facilidad por el modelo, el cual clasifica el 90,0 % de manera correcta.

La expresión dada por el modelo para clasificar a los estudiantes se muestra a continuación:

$$P(Grupos\_Segundo = 1) = \frac{1}{1 + e^{(21{,}885 - 0{,}675 \cdot Nota\_Sele - 2{,}091 \cdot Promedio\_Primero)}}$$

Si $P(Grupos\_Segundo = 1) \geq 0{,}5$ se clasifica al estudiante en el grupo 1, que son aquellos alumnos por encima de la media en segundo.

6.6.3. Predicción de los alumnos que repetirán alguna asignatura en segundo

Se procede a construir el modelo para la variable "Repiten_Segundo_S_N", que es la última que falta para completar el análisis de las variables que definen el comportamiento del alumno en segundo curso. El filtrado de datos en este caso consiste en seleccionar únicamente a aquellos alumnos que han llegado a segundo (hayan completado el curso o no). Las características principales de los datos para construir el modelo son las siguientes:

- Datos empleados para la creación del modelo = 1324 (90,3 %)
- Datos empleados para la validación del modelo = 142 (9,7 %)
- Datos totales = 1466

Codificación de la variable dependiente:

- Repiten_Segundo_S_N = 0 $\rightarrow$ Alumnos que no repiten ninguna asignatura de segundo.
- Repiten_Segundo_S_N = 1 $\rightarrow$ Alumnos que repiten una o más asignaturas en segundo.

Variables preseleccionadas para la creación del modelo:

- Nota_Sele
- Promedio_Primero
- Grupos_Primero
- Repetidas_Primero
- Alpha_Primero

- Completan_Primero_Al_Ritmo

De los alumnos que han llegado a segundo, cerca de un 70 % repite alguna asignatura, por lo que el grupo con mayor frecuencia es el Repiten_Segundo_S_N = 1. La primera clasificación realizada en el bloque 0 consiste en etiquetar todos los casos en dicho grupo:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Repiten_Segundo_S_N		Porcentaje correcto	Repiten_Segundo_S_N		Porcentaje correcto
	Observado		0	1		0	1	
Paso 0	Repiten_Segundo_S_N	0	0	399	,0	0	44	,0
		1	0	925	100,0	0	98	100,0
	Porcentaje global				69,9			69,0

Tabla 6.76. Clasificación realizada en el bloque 0

El modelo se ha construido en dos pasos, introduciendo finalmente dos variables Promedio_Primero en el primer paso, y Alpha_Primero en el segundo paso:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Promedio_Primero	-2,045	,125	268,383	1	,000	,129
	Constante	14,074	,814	298,744	1	,000	1294888
Paso 2[b]	Promedio_Primero	-1,775	,139	163,449	1	,000	,170
	Alpha_Primero	-2,230	,602	13,722	1	,000	,108
	Constante	14,290	,814	308,216	1	,000	1606940

a. Variable(s) introducida(s) en el paso 1: Promedio_Primero.

b. Variable(s) introducida(s) en el paso 2: Alpha_Primero.

Tabla 6.77. Variables introducidas en la ecuación con coeficientes B.

Observando el parámetro B, por un lado se puede apreciar como aquellos alumnos con un promedio alto en primero tienen tendencia a pertenecer al grupo 0 (alumnos que no repiten en segundo). Por otro lado, aquellos alumnos con un parámetro Alpha en primero, que son aquellos que no repiten o repiten pocas asignaturas en primero, también tienen más probabilidad de pertenecer al grupo 0 en segundo.

Siguiendo los mismos pasos que en el modelo anterior, se muestra ahora el valor de la R Cuadrado de Nagelkerke y el resultado de la prueba de Hosmer y Lemeshow:

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1188,687[a]	,278	,394
2	1174,337[a]	,286	,405

Tabla 6.78. Valor de la R Cuadrado de Nagelkerke

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	17,723	8	,023
2	6,151	8	,630

Tabla 6.79. Sigma de la prueba de Hosmer y Lemeshow

Dado que el valor de la sigma es 0,630, y es mayor al nivel de significación de 0,05, se pude afirmar que el modelo se ajusta de manera significativa a los datos reales. EL porcentaje de varianza explicado es el 40,5 %, tal y como se indica con el valor de la R Cuadrado de Nagelkerke.

La clasificación realizada por el modelo, tanto para el subconjunto de entrenamiento como para el subconjunto de validación, es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Repiten_Segundo_S_N		Porcentaje correcto	Repiten_Segundo_S_N		Porcentaje correcto
	Observado		0	1		0	1	
Paso 2	Repiten_Segundo_S_N	0	221	178	55,4	24	20	54,5
		1	67	858	92,8	8	90	91,8
	Porcentaje global				81,5			80,3

Tabla 6.80. Clasificación realizada por el modelo generado.

Se consigue un porcentaje de acierto de más del 90 % para el grupo 1, que son aquellos estudiantes que repiten alguna asignatura. Para el grupo 0, el porcentaje de acierto está alrededor del 55 %. En total, el porcentaje de alumnos clasificados con éxito es del 80,3 %.

La ecuación generada por el modelo para realizar la clasificación de estudiantes en grupos es la siguiente:

$$P(\text{Re } piten\_Segundo\_S\_N=1)=\frac{1}{1+e^{(-14,290+2,230\cdot Alpha\_Primero+1,775\cdot Promedio\_Primero)}}$$

Si $P(\text{Re } piten\_Segundo\_S\_N=1)$ es igual o superior al punto de corte 0,5, se clasifica al estudiante en el grupo 1, que son aquellos alumnos que repiten en segundo.

6.7. Modelo predictivo para los alumnos en tercero y cuarto

En este apartado se muestra un resumen de las características más relevantes de los modelos generados para tercero y cuarto. No obstante, para evitar realizar una memoria demasiado extensa, los resultados completos se han incluido en el anexo.

6.7.1. Modelos para el tercer curso

En primer lugar, se muestra el modelo generado para pronosticar si el alumno completará el tercer curso en el tercer año de carrera (Completan_Tercero_Al_Ritmo). Para la creación del modelo, se han considerado únicamente los alumnos que han ingresado en los años 2010 y 2011, que son los únicos que pueden haber completado tercero yendo a curso por año. En total, se dispone de 689 casos para la creación del modelo y de 126 para la validación del mismo.

El modelo generado es el siguiente para determinar la probabilidad de que el alumno complete tercero al ritmo marcado $P(Completan\_Tercero\_Al\_Ritmo=1)$ es la siguiente:

$$P(Completan\_Tercero\_Al\_Ritmo=1)=\frac{1}{1+e^{\begin{array}{l}(10,904-1,661\cdot Completan\_Segundo\_Al\_Ritmo+\\+0,312\cdot \text{Re } petidas\_Segundo+\\(-1,334)\cdot Promedio\_Segundo+\\(-1,201)\cdot Completan\_Primero\_Al\_Ritmo)\end{array}}}$$

Si dicha probabilidad es igual o superior a 0,5, se clasifica al estudiante en el grupo 1, que son aquellos estudiantes que completan tercero en tres años. La tabla de clasificación empleando el modelo se muestra a continuación:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b,c]		
	Observado		Completan_tercero_Al_Ritmo 0	Completan_tercero_Al_Ritmo 1	Porcentaje correcto	Completan_tercero_Al_Ritmo 0	Completan_tercero_Al_Ritmo 1	Porcentaje correcto
Paso 4	Completan_tercero_Al_Ritmo	0	472	47	90,9	68	9	88,3
		1	34	136	80,0	7	28	80,0
	Porcentaje global				88,2			85,7

Tabla 6.81. Tabla de clasificación empleando el modelo.

Se observa que un 85,7 % de los casos se clasifican con éxito para el subconjunto de validación.

El siguiente modelo que se muestra es el confeccionado para determinar si el alumno estará por encima o por debajo de la media de tercero. Para su creación, se han empleado únicamente aquellos estudiantes que han completado tercero, que son un total de 372, de los cuales 275 se han empleado para la creación del modelo y 97 para su validación.

La expresión que determina la probabilidad de que un alumno esté por encima de la media en tercero (Grupos_Tercero = 1) es la siguiente:

$$P(Grupos\_Tercero = 1) = \frac{1}{1 + e^{(19{,}164 - 0{,}858 \cdot Completan\_Segundo\_Al\_Ritmo + (-2{,}809) \cdot Promedio\_Segundo)}}$$

El punto de corte para la clasificación es 0,5. Si $P(Grupos\_Tercero = 1) \geq 0{,}5$ se clasifica al estudiante en el grupo 1, que son aquellos que están por encima de la media en tercero. La tabla de clasificación ofrecida por el programa es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
	Observado		Grupos_Tercero 0	Grupos_Tercero 1	Porcentaje correcto	Grupos_Tercero 0	Grupos_Tercero 1	Porcentaje correcto
Paso 1	Grupos_Tercero	0	125	25	83,3	52	8	86,7
		1	36	89	71,2	11	26	70,3
	Porcentaje global				77,8			80,4
Paso 2	Grupos_Tercero	0	123	27	82,0	51	9	85,0
		1	30	95	76,0	6	31	83,8
	Porcentaje global				79,3			84,5

Tabla 6.82. Tabla de clasificación empleando el modelo.

Se observa como un 84,5 % de los casos son clasificados con éxito para el subconjunto de validación.

El último modelo del tercer curso generado permite pronosticar si el alumno repetirá alguna asignatura o no en tercero. Para su creación, se han considerado únicamente aquellos alumnos que han llegado al tercer curso. En total, se disponen de 1000 registros, de los cuales 866 son utilizados para crear el modelo, y 134 para validarlo.

La ecuación dada por el modelo para determinar la probabilidad de que el alumno repita en tercero (Repiten_Tercero_S_N = 1) es la siguiente:

$$P(\text{Re}\,piten\_Tercero\_S\_N=1)=\frac{1}{1+e^{(-10,700-0,585\cdot \text{Re}\,piten\_Segundo\_S\_N+1,103\cdot Grupos\_Segundo+1,219\cdot \text{Promedio}\_Segundo+0,446\cdot \text{Promedio}\_\text{Primero})}}$$

Si $P(\text{Re}\,piten\_Tercero\_S\_N=1)\geq 0,5$ se pronostica que el alumno repetirá en tercero. La tabla de clasificación ofrecida por el programa es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Repiten_Tercero_S_N		Porcentaje correcto	Repiten_Tercero_S_N		Porcentaje correcto
	Observado		0	1		0	1	
Paso 4	Repiten_Tercero_S_N	0	231	125	64,9	37	21	63,8
		1	44	466	91,4	13	63	82,9
	Porcentaje global				80,5			74,6

Tabla 6.83. Tabla de clasificación empleando el modelo

El 74,6 % de los casos se clasifican con éxito empleando el modelo generado en el subconjunto de validación.

6.7.2. Modelos para el cuarto curso

El primer modelo mostrado es el que sirve para pronosticar si el alumno concluirá los estudios en cuatro años o se demorará (Completan_Cuarto_Al_Ritmo). Para la creación del modelo, se han considerado únicamente los alumnos que han ingresado en el año 2010, que son los únicos que pueden haber completado la titulación yendo a curso por año. En total, se dispone de 359 casos para la creación del modelo y de 64 para su validación.

La expresión que determina la probabilidad de que un alumno complete la titulación en cuatro años basándose en los resultados obtenidos en primero y en segundo es la siguiente:

$$P(Comple\tan\_Cuarto\_Al\_Ritmo=1)=\frac{1}{1+e^{\begin{array}{l}(10{,}730-1{,}323\cdot Comple\tan\_Segundo\_Al\_Ritmo+\\(-11{,}681)\cdot Alpha\_Segundo+\\(-1{,}197)\cdot \mathrm{Re}piten\_Segundo\_S\_N+\\+0{,}391\cdot \mathrm{Re}petidas\_\mathrm{P}rimero+0{,}566\cdot Grupos\_\mathrm{P}rimero)\end{array}}}$$

Como en modelos anteriores, el punto de corte está en 0,5. Si la $P(Comple\tan\_Cuarto\_Al\_Ritmo=1)$ es igual o superior a dicho valor, se clasifica al alumno en el grupo 1, que son aquellos que completan la titulación en cuatro años. Los resultados obtenidos aplicando el modelo son los siguientes:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b,c]		
			Completan_Cuarto_Al_Ritmo		Porcentaje correcto	Completan_Cuarto_Al_Ritmo		Porcentaje correcto
	Observado		0	1		0	1	
Paso 5	Completan_Cuarto_Al_Ritmo	0	189	29	86,7	24	3	88,9
		1	30	111	78,7	3	25	89,3
	Porcentaje global				83,6			89,1

Tabla 6.84. Tabla de clasificación empleando el modelo generado

Para el subconjunto de validación, el porcentaje total de acierto es del 89,1 %.

A continuación se muestran las características principales del modelo para pronosticar si el alumno estará por encima o por debajo de la media de cuarto (Grupos_Cuarto). Para crear este modelo, se han empleado 129 casos para el análisis, y 41 para la validación. La cantidad total de datos es pequeña debido a que hay pocos estudiantes que hayan terminado la titulación en 2014.

La ecuación generada por el modelo para la clasificación de casos es la siguiente:

$$P(Grupos\_Cuarto=1)=\frac{1}{1+e^{(21{,}107-3{,}248\cdot \mathrm{P}romedio\_Tercero)}}$$

El punto de corte para la clasificación es 0,5. Si $P(Grupos\_Cuarto=1)\geq 0{,}5$ se clasifica al estudiante en el grupo 1, que son aquellos que están por encima de la media en cuarto. La tabla de clasificación ofrecida por el programa es la siguiente:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b]		
			Grupos_Cuarto		Porcentaje correcto	Grupos_Cuarto		Porcentaje correcto
	Observado		0	1		0	1	
Paso 1	Grupos_Cuarto	0	54	12	81,8	16	5	76,2
		1	14	49	77,8	4	16	80,0
	Porcentaje global				79,8			78,0

Tabla 6.85. Tabla de clasificación empleando el modelo generado

Se observa como un 78,0 % de los casos son clasificados con éxito para el subconjunto de validación.

Por último, se desea pronosticar si el alumno repetirá o no en cuarto (Repiten_Cuarto_S_N). Para elaborar el modelo se utilizan únicamente aquellos datos pertenecientes a alumnos que han llegado a cuarto. Se utilizan 461 para la construcción del modelo, y 101 para su validación.

La expresión dada por el modelo para calcular la probabilidad de que un alumno repita en cuarto es la siguiente:

$$P(\text{Re}\ piten\_Cuarto\_S\_N=1)=\frac{1}{1+e^{\begin{array}{l}(-0{,}544+2{,}890\cdot Alpha\_Tercero+\\(-1{,}052)\cdot \text{Re}\ piten\_Tercero\_S\_N+\\+1{,}209\cdot Grupos\_Tercero-0{,}932\cdot \text{Re}\ piten\_Segundo\_S\_N)\end{array}}}$$

Si $P(\text{Re}\ piten\_Cuarto\_S\_N=1)$ es igual o superior a 0,5, se clasifica al estudiante en el grupo 1, que es el de los estudiantes que repiten alguna asignatura en cuarto. Siguiendo este criterio de clasificación, los resultados obtenidos para el subconjunto de validación son los siguientes:

			Pronosticado					
			Casos seleccionados[a]			Casos no seleccionados[b,c]		
			Repiten_Cuarto_S_N		Porcentaje correcto	Repiten_Cuarto_S_N		Porcentaje correcto
	Observado		0	1		0	1	
Paso 4	Repiten_Cuarto_S_N	0	233	65	78,2	56	9	86,2
		1	45	118	72,4	12	23	65,7
	Porcentaje global				76,1			79,0

Tabla 6.86. Tabla de clasificación empleando el modelo generado

Se predice con un 79,0 % de acierto si un alumno repetirá o no en cuarto.

6.8. Predicción de las notas de las distintas asignaturas

En este apartado se muestran los modelos construidos para predecir las notas de las distintas asignaturas. Debido a que tanto las variables dependientes como las independientes son cuantitativas, y todas ellas están en una misma escala que va del 0,0 al 10,0, la técnica empleada para la predicción de notas es la regresión lineal. Por lo tanto, lo que se desea construir es una combinación lineal con las notas de asignaturas ya realizadas, para estimar la nota de asignaturas futuras. Del mismo modo que sucede con la regresión logística, existen distintos métodos para la selección de variables que participan en el modelo. En este modelo se ha empleado el método hacia adelante, que consiste en partir de un modelo formado únicamente por una constante, e ir añadiendo las distintas variables en función de su significación. Para emplear este método, previamente se debe haber preseleccionado un abanico de variables, que son las que se introducen o no una vez contrastadas por el propio modelo. En este caso, el abanico de variables preseleccionadas está formado por todas aquellas asignaturas que el alumno ha cursado previamente. Por ejemplo, si se desea predecir la nota de Dinámica de Sistemas (segundo cuatrimestre del segundo curso), las variables preseleccionadas serán las notas de las asignaturas del primer curso, y del primer cuatrimestre de segundo.

Debido al gran número de asignaturas de la titulación, a continuación se muestra a modo de ejemplo los resultados completos obtenidos para una de las asignaturas (Estadística), pero para el resto se mostrará únicamente una síntesis de los resultados. No obstante, los resultados completos de todas las asignaturas se pueden encontrar en el anexo.

Predicción de la nota de Estadística

Estadística es una asignatura impartida en el segundo cuatrimestre del segundo curso, por lo que se disponen de las notas de los alumnos en el primer curso y en el primer cuatrimestre del segundo curso. Del mismo modo que se ha realizado en las regresiones logísticas, se debe crear un subconjunto de validación que no participe en la creación del modelo para medir su precisión.

- Variable dependiente: Estadística
- Variables preseleccionadas:
  - Álgebra lineal
  - Cálculo I
  - Mecánica Fundamental

- Química I
- Fundamentos de Informática
- Geometría
- Cálculo II
- Termodinámica Fundamental
- Química II
- Expresión Gráfica
- Electromagnetismo
- Métodos Numéricos
- Materiales
- Ecuaciones Diferenciales
- Informática
- Mecánica

- Número de filas del archivo = 1820
- Subconjunto de validación = 148

El primer cuadro importante que ofrece el programa SPSS es el siguiente:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
	Aleatorio <= 85 (Seleccionado)	Cambio en R cuadrado	Cambio en F	gl1
1	,484(a)	,234	,233	1,21690
2	,538(b)	,290	,288	1,17234
3	,576(c)	,332	,330	1,13762
4	,595(d)	,354	,351	1,11959
5	,607(e)	,369	,365	1,10685
6	,616(f)	,379	,375	1,09846
7	,619(g)	,384	,379	1,09512

a Variables predictoras: (Constante), Ecuaciones_Diferenciales
b Variables predictoras: (Constante), Ecuaciones_Diferenciales, Mecanica
c Variables predictoras: (Constante), Ecuaciones_Diferenciales, Mecanica, Informatica
d Variables predictoras: (Constante), Ecuaciones_Diferenciales, Mecanica, Informatica, Algebra_Lineal
e Variables predictoras: (Constante), Ecuaciones_Diferenciales, Mecanica, Informatica, Algebra_Lineal, Metodos_Numericos
f Variables predictoras: (Constante), Ecuaciones_Diferenciales, Mecanica, Informatica, Algebra_Lineal, Metodos_Numericos, Fundamentos_Informatica
g Variables predictoras: (Constante), Ecuaciones_Diferenciales, Mecanica, Informatica, Algebra_Lineal, Metodos_Numericos, Fundamentos_Informatica, Materiales

Tabla 6.87. Variables introducidas en cada paso con los errores típicos

Los parámetros mostrados se estructuran por pasos, en cada uno de los cuales se introduce una nueva variable al modelo. De todas las variables preseleccionadas, las introducidas en este caso son las siguientes:

- **Paso 1:** Ecuaciones Diferenciales
- **Paso 2:** Mecánica
- **Paso 3:** Informática
- **Paso 4:** Álgebra Lineal
- **Paso 5:** Métodos Numéricos
- **Paso 6:** Fundamentos de Informática
- **Paso 7:** Materiales

Para cada uno de los pasos, se muestra también el error típico de la estimación, que es el valor promedio de los errores y que adopta un valor de 1,095. Otro valor interesante es el R Cuadrado, que da una idea de en qué medida influye el modelo sobre la variable dependiente. En el caso de Estadística, se puede afirmar que el modelo final influye en un

38,4 % en el valor de la variable dependiente. Para observar si el modelo es significativo o no se debe observar el valor sigma de la siguiente tabla:

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
	Regresión	666,749	7	95,250	79,422	,000(g)
	Residual	1070,960	893	1,199		
	Total	1737,709	900			

Tabla 6.88. Significación del modelo generado

Dado que el valor de sigma es 0,000 y es inferior al nivel de significación de 0,05, se puede afirmar que el modelo sí que es significativo.

Se comprueba también que la hipótesis de que los residuos tipificados siguen una distribución normal con media igual a 0 es cierta representándolos en un histograma:

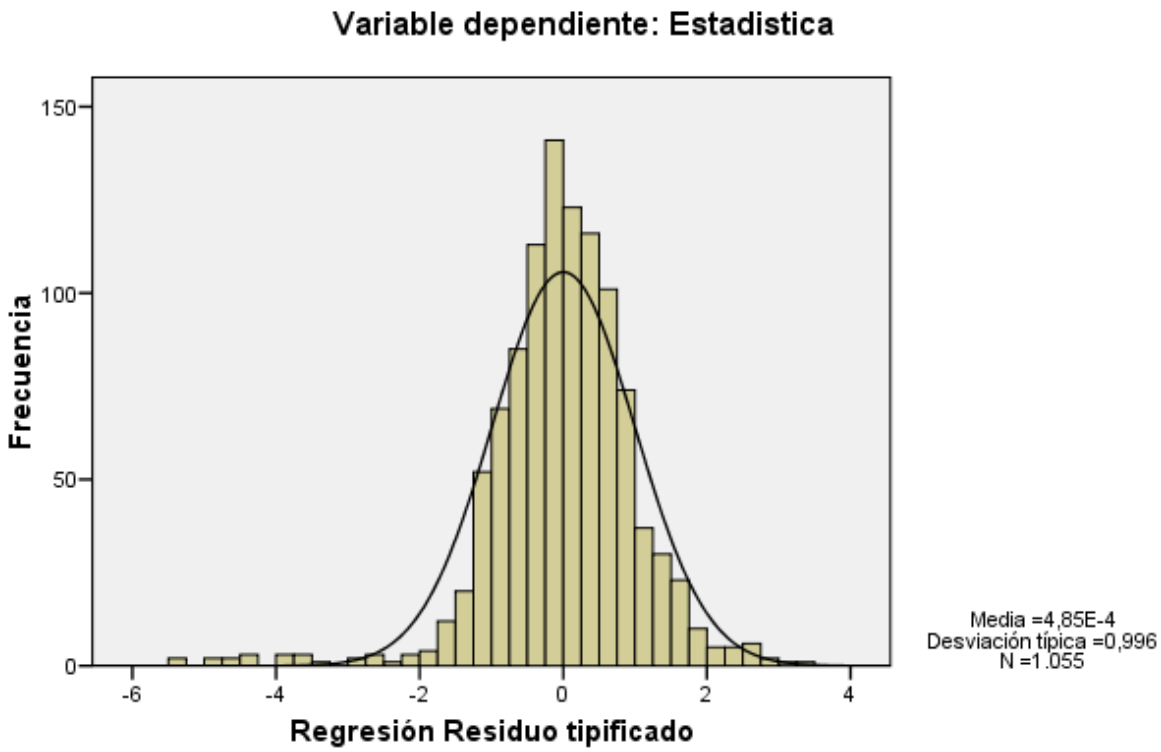

Figura 6.17. Histograma de los residuos tipificados del modelo

La siguiente tabla que ofrece el programa SPSS y quizás la más importante es la siguiente:

Coeficientes

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Error típ.	Beta	Orden cero	Parcial	Semiparcial	B	Error típ.
7	(Constante)	2,106	,187		11,267	,000			
	Ecuaciones_Diferenciales	,123	,029	,152	4,315	,000	,484	,143	,113
	Mecanica	,111	,022	,157	5,094	,000	,417	,168	,134
	Informatica	,108	,025	,143	4,424	,000	,436	,146	,116
	Algebra_Lineal	,097	,028	,110	3,478	,001	,405	,116	,091
	Metodos_Numericos	,114	,028	,131	4,038	,000	,437	,134	,106
	Fundamentos_Informatica	,064	,018	,113	3,596	,000	,394	,119	,094
	Materiales	,083	,033	,085	2,544	,011	,428	,085	,067

Tabla 6.89. Coeficientes proporcionados por la regresión lineal

El parámetro B de la tabla son los coeficientes de cada una de las variables de la regresión lineal. También se muestra la significación de cada una de las variables con el valor de sigma (si es inferior a 0,05 la variable es significativa), y la correlación semiparcial con la variable dependiente. Construyendo el modelo que muestra la tabla, la nota de Estadística viene dada por la siguiente expresión:

$$Nota\_Estadistica = 2{,}106 + 0{,}123 \cdot Nota\_Ec.Diferenciales + 0{,}111 \cdot Nota\_Mecanica +$$
$$+ 0{,}108 \cdot Nota\_Informatica + 0{,}097 \cdot Nota\_Alg.Lineal +$$
$$+ 0{,}114 \cdot Nota\_Met.Numericos + 0{,}064 \cdot Nota\_Fund.Informatica +$$
$$+ 0{,}083 \cdot Nota\_Materiales$$

Se ha observado que el error típico de la regresión lineal construida es de 1,095. No obstante, este valor puede verse muy incrementado por ejemplo, en alumnos que hayan abandonado la asignatura a medio curso y cuya nota sea excesivamente baja. Para reducir los efectos de los casos extremos, se realiza una segunda comprobación del modelo codificando las notas en distintos grupos:

- Notas menores a 5 → Grupo 1 (Suspensos)
- Entre 5 y 7 (con el 5 incluido) → Grupo 2 (Aprobados)
- Entre 7 y 9 (con el 7 incluido) → Grupo 3 (Notables)
- Notas iguales o superiores a 9 → Grupo 4 (Excelentes)

El método es el siguiente: para cada uno de los casos del subconjunto de validación, se colocará su nota real en su correspondiente grupo. Lo mismo se realizará para la nota estimada por la regresión lineal. Con las notas codificadas en los grupos, se observará qué porcentaje de acierto proporciona la expresión dada por la regresión lineal. Los resultados son los siguientes:

Estadistica	
Aciertos	96
Total	148
% Acierto	64,86%

Tabla 6.90. Porcentaje de acierto en las cualificaciones

Del total de 148 datos del subconjunto de validación, 96 han sido clasificados con una cualificación correcta, lo que representa un 64,86 % de acierto. Se ha querido comprobar también la efectividad del modelo a la hora de determinar si el alumno aprobará o suspenderá la asignatura. Para ello, en este caso las notas se han codificado en dos grupos:

- Notas menores a 5 → Grupo 1 (Suspensos)
- Notas iguales o superiores a 5 → Grupo 2 (Aprobados)

Con la nueva codificación de los casos, los resultados obtenidos se muestran en la siguiente tabla:

Aciertos	**131**
Total	148
% Acierto	88,51%

Tabla 6.91. Porcentaje de acierto en aprobados y suspendidos

Se puede apreciar como para el caso de Estadística, de los 148 casos del subconjunto de validación, el 88,51 % han sido clasificados con éxito. Por lo tanto, se concluye que el modelo generado para dicha asignatura ofrece unos buenos resultados.

6.8.1. Modelos de las distintas asignaturas

En este apartado se muestran las distintas ecuaciones obtenidas con las regresiones lineales para predecir la nota de las distintas asignaturas.

Geometría

$$Nota\_Geometria = 1{,}14 + 0{,}149 \cdot Nota\_Alg.Lineal + 0{,}25 \cdot Nota\_CalculoI + \\ + 0{,}288 \cdot Nota\_Mecanica.Fund. + 0{,}095 \cdot Nota\_Fund.Informatica$$

Cálculo II

$$Nota\_CalculoII = 0{,}984 + 0{,}258 \cdot Nota\_Alg.Lineal + 0{,}284 \cdot Nota\_CalculoI + \\ + 0{,}231 \cdot Nota\_Mecanica.Fund.$$

Termodinámica Fundamental

$$Nota\_Termodinamica.Fund = 0{,}485 + 0{,}434 \cdot Nota\_Alg.Lineal + \\ + 0{,}185 \cdot Nota\_CalculoI + 0{,}198 \cdot Nota\_Mecanica.Fund.$$

Química II

$$Nota\_QuimicaII = 1{,}717 + 0{,}254 \cdot Nota\_Alg.Lineal + 0{,}235 \cdot Nota\_CalculoI + \\ + 0{,}119 \cdot Nota\_Mecanica.Fund. + 0{,}072 \cdot Nota\_QuimicaI + \\ + 0{,}047 \cdot Nota\_Fund.Informatica$$

Expresión Gráfica

$$Nota\_Exp.Grafica = 1{,}242 + 0{,}182 \cdot Nota\_Alg.Lineal + 0{,}169 \cdot Nota\_CalculoI + \\ + 0{,}175 \cdot Nota\_Mecanica.Fund. + 0{,}237 \cdot Nota\_Fund.Informatica$$

Electromagnetismo

$$Nota\_Electromagnetismo = 0{,}191 + 0{,}108 \cdot Nota\_Alg.Lineal + \\ + 0{,}124 \cdot Nota\_QuimicaI + 0{,}222 \cdot Nota\_Mecanica.Fund. + \\ (-0{,}047) \cdot Nota\_Fund.Informatica + 0{,}091 \cdot Nota\_CalculoII + \\ + 0{,}181 \cdot Nota\_Termodinamica.Fund + 0{,}178 \cdot Nota\_QuimicaII$$

Métodos Numéricos

$$Nota\_MetodosNumeri\cos = 1{,}619 + 0{,}083 \cdot Nota\_Alg.Lineal +$$
$$+ 0{,}095 \cdot Nota\_Mecanica.Fund. + 0{,}075 \cdot Nota\_Fund.Informatica +$$
$$+ 0{,}100 \cdot Nota\_CalculoII + 0{,}085 \cdot Nota\_Ter\,\mathrm{mod}\,inamica.Fund +$$
$$+ 0{,}216 \cdot Nota\_QuimicaII + 0{,}093 \cdot Nota\_Geometria$$

Materiales

$$Nota\_Materiales = 1{,}065 + 0{,}197 \cdot Nota\_Mecanica.Fund + 0{,}055 \cdot Nota\_QuimicaI +$$
$$+ 0{,}051 \cdot Nota\_Funf.Informatica + 0{,}082 \cdot Nota\_CalculoII +$$
$$+ 0{,}088 \cdot Nota\_Ter\,\mathrm{mod}\,inamicaFund. + 0{,}239 \cdot Nota\_QuimicaII$$

Ecuaciones Diferenciales

$$Nota\_Ec.Diferenciales = 0{,}19 + 0{,}196 \cdot Nota\_CalculoI +$$
$$+ 0{,}199 \cdot Nota\_Mecanica.Fund + 0{,}125 \cdot Nota\_CalculoII +$$
$$+ 0{,}092 \cdot Nota\_Ter\,\mathrm{mod}\,inamica.Fund + 0{,}268 \cdot Nota\_QuimicaII$$

Informática

$$Nota\_Informatica = 1{,}163 + 0{,}123 \cdot Nota\_Mecanica.Fund + 0{,}108 \cdot Nota\_QuimicaI +$$
$$+ 0{,}161 \cdot Nota\_Fund.Informatica + 0{,}203 \cdot Nota\_QuimicaII +$$
$$+ 0{,}209 \cdot Nota\_ExpresionGrafica$$

Mecánica

$$Nota\_Mecanica = -0{,}772 + 0{,}122 \cdot Nota\_CalculoI + 0{,}191 \cdot Nota\_Mecnica.Fund +$$
$$+ 0{,}325 \cdot Nota\_Ter\,\mathrm{mod}\,inamica.Fund. + 0{,}125 \cdot Nota\_QuimicaII +$$
$$+ 0{,}119 \cdot Nota\_Exp.Grafica$$

Economía y Empresa

$$Nota\_EconomiaEmpresa = 2{,}978 + 0{,}115 \cdot Nota\_Mecanica.Fund. +$$
$$+ 0{,}096 \cdot Nota\_Fund.Informatica + 0{,}084 \cdot Nota\_MetodosNumeri\cos +$$
$$+ 0{,}055 \cdot Nota\_Mecanica + 0{,}053 \cdot Nota\_QuimicaII + 0{,}069 \cdot Nota\_Materiales +$$
$$+ 0{,}072 \cdot Nota\_Geometria + 0{,}055 \cdot Nota\_EcuacionesDiferenciales$$

Estadística

$$Nota\_Estadistica = 2{,}106 + 0{,}123 \cdot Nota\_Ec.Diferenciales + 0{,}111 \cdot Nota\_Mecanica +$$
$$+ 0{,}108 \cdot Nota\_Informatica + 0{,}097 \cdot Nota\_Alg.Lineal + 0{,}114 \cdot Nota\_Met.Numericos +$$
$$+ 0{,}064 \cdot Nota\_Fund.Informatica + 0{,}083 \cdot Nota\_Materiales$$

Dinámica de Sistemas

$$Nota\_DinamicaSistemas = 0{,}525 + 0{,}162 \cdot Nota\_EcuacionesDiferenciales +$$
$$+ 0{,}134 \cdot Nota\_Mecanica + 0{,}124 \cdot Nota\_AlgebraLineal +$$
$$+ 0{,}153 \cdot Nota\_Metodos.Numericos + 0{,}150 \cdot Nota\_Geometria +$$
$$+ 0{,}104 \cdot Nota\_Electromagnetismo + 0{,}097 \cdot Nota\_Materiales +$$
$$+ 0{,}077 \cdot Nota\_Termodinamica.Fund$$

Proyecto I

$$Nota\_ProyectoI = 7{,}498 - 0{,}09 \cdot Nota\_Ec.Diferenciales + 0{,}099 \cdot Nota\_Informatica +$$
$$+ 0{,}054 \cdot Nota\_Fund.Informatica + 0{,}082 \cdot Nota\_MetodosNumericos$$

Teoría de Máquinas y Mecanismos

$$Nota\_TeoríaMaquinasMecanismos = 0{,}369 + 0{,}228 \cdot Nota\_Mecanica +$$
$$+ 0{,}166 \cdot Nota\_EcuacionesDif. + 0{,}167 \cdot Nota\_ExpresionGrafica +$$
$$+ 0{,}162 \cdot Nota\_Electromagnetismo + 0{,}152 \cdot Nota\_MetodosNumericos +$$
$$+ 0{,}064 \cdot Nota\_Fund.Informatica + 0{,}142 \cdot Nota\_Mecanica.Fund +$$
$$(-0{,}093) \cdot Nota\_CalculoI$$

Tecnología del Medio Ambiente y Sostenibilidad

$$Nota\_Tec.MedioAmb.y.Sost = -0{,}073 + 0{,}256 \cdot Nota\_Din.Sistemas +$$
$$+ 0{,}194 \cdot Nota\_Materiales + 0{,}114 \cdot Nota\_Alg.Lineal + 0{,}095 \cdot Nota\_QuimicaII +$$
$$0{,}084 \cdot Nota\_Mecanica + 0{,}094 \cdot Nota\_QuimicaI + 0{,}082 \cdot Nota\_Met.Numericos +$$
$$+ 0{,}055 \cdot Nota\_Exp.Grafica$$

Termodinámica

$$Nota\_Termodinamica = 0{,}708 + 0{,}287 \cdot Nota\_Din.Sistemas + 0{,}273 \cdot Nota\_Estadistica + \\ + 0{,}127 \cdot Nota\_TeoriaMaquinas + 0{,}134 \cdot Nota\_QuimicaII + 0{,}072 \cdot Nota\_Informatica + \\ (-0{,}095) \cdot Nota\_Geometria + 0{,}06 \cdot Nota\_Electromagnetismo$$

Electrotecnia

$$Nota\_Electrotecnia = -0{,}796 + 0{,}266 \cdot Nota\_Din.Sistemas + 0{,}260 \cdot Nota\_Estadistica + \\ + 0{,}151 \cdot Nota\_Mecanica + 0{,}098 \cdot Nota\_QuimicaII + 0{,}108 \cdot Nota\_TeoriaMaquinas + \\ + 0{,}104 \cdot Nota\_CalculoII$$

Mecánica de Medios Continuos

$$Nota\_Mec.Medios.Cont = 0{,}501 + 0{,}185 \cdot Nota\_Teoria.Maquinas + \\ + 0{,}142 \cdot Nota\_Din.Sistemas + 0{,}15 \cdot Nota\_Electromagnetismo + 0{,}123 \cdot Nota\_Materiales + \\ + 0{,}088 \cdot Nota\_Met.Numericos + 0{,}095 \cdot Nota\_Geometria + 0{,}076 \cdot Nota\_Ec..Diferenciales$$

Técnicas Estadísticas para la Calidad

$$Nota\_Tec.Estadisticas.Calidad = 2{,}129 + 0{,}184 \cdot Nota\_Din.Sistemas + \\ + 0{,}177 \cdot Nota\_Estadistica + 0{,}14 \cdot Nota\_Econ.Empresa + 0{,}096 \cdot Nota\_Proyecto I + \\ + 0{,}033 \cdot Nota\_Fund.Informatica - 0{,}097 \cdot Nota\_Electromagnetismo + \\ + 0{,}067 \cdot Nota\_QuimicaII + 0{,}049 \cdot Nota\_Mecanica - 0{,}074 \cdot Nota\_Geometria + \\ + 0{,}054 \cdot Nota\_Informatica$$

Tecnología de Selección de Materiales

$$Nota\_Tec.Selec.Materiales = 0{,}823 + 0{,}183 \cdot Nota\_Din.Sistemas + \\ + 0{,}211 \cdot Nota\_Materiales + 0{,}087 \cdot Nota\_Teoria.Maquinas + 0{,}155 \cdot Nota\_Proyecto I + \\ + 0{,}118 \cdot Nota\_QuimicaII - 0{,}096 \cdot Nota\_CalculoII + 0{,}095 \cdot Nota\_Estadistica$$

Mecánica de Fluidos

$$Nota\_Mec.Fluidos = -0{,}343 + 0{,}169 \cdot Nota\_Termodinamica + 0{,}132 \cdot Nota\_Electrotecnia + \\ + 0{,}136 \cdot Nota\_Materiales + 0{,}112 \cdot Nota\_Tec.Medio.Ambiente + \\ + 0{,}095 \cdot Nota\_Termodinamica.Fund. - 0{,}09 \cdot Nota\_Fund.Informatica + \\ + 0{,}108 \cdot Nota\_Mec.Medios.Cont. + 0{,}143 \cdot Nota\_Proyecto I + 0{,}1 \cdot Nota\_Mecanica.Fund.$$

Organización y Gestión

$$Nota\_Org.y.Gestion = 1{,}373 + 0{,}164 \cdot Nota\_Ter\,\text{mod}\,inamica + 0{,}168 \cdot Nota\_Informatica +$$
$$+ 0{,}081 \cdot Nota\_Tec.Medio.Ambiente + 0{,}132 \cdot Nota\_Tec.Selec.Materiales +$$
$$+ 0{,}106 \cdot Nota\_Mec.Medios.Cont. - 0{,}08 \cdot Nota\_QuimicaI +$$
$$+ 0{,}088 \cdot Nota\_Economia.y.Empresa.$$

Resistencia de Materiales

$$Nota\_\text{Re}\,sist.Materiales = 0{,}557 + 0{,}19 \cdot Nota\_Ec.Diferenciales +$$
$$+ 0{,}119 \cdot Nota\_Tec.Medio.Ambiente + 0{,}106 \cdot Nota\_Electrotecnia +$$
$$+ 0{,}15 \cdot Nota\_Tec.Selec.Materiales + 0{,}123 \cdot Nota\_Ter\,\text{mod}\,inamica +$$
$$0{,}106 \cdot Nota\_Mecanica.Fund. + 0{,}074 \cdot Nota\_Teoria.Maq.y.Mecanismos +$$
$$+ 0{,}08 \cdot Nota\_Din.Sistemas$$

Proyecto II

$$Nota\_\text{Pr}\,oyectoII = 6{,}962 + 0{,}045 \cdot Nota\_Exp.Grafica + 0{,}116 \cdot Nota\_Tec.Estad.Calidad +$$
$$+ 0{,}075 \cdot Nota\_Informatica + 0{,}064 \cdot Nota\_Teoria.Maq.y.Mecanismos +$$
$$(-0{,}055) \cdot Nota\_QuimicaI$$

Máquinas Eléctricas

$$Nota\_Maquinas.Electricas = 3{,}398 + 0{,}146 \cdot Nota\_Electrotecnia +$$
$$+ 0{,}233 \cdot Nota\_Tec.Estadist.Calidad + 0{,}158 \cdot Nota\_Ter\,\text{mod}\,inamica +$$
$$(-0{,}181) \cdot Nota\_Informatica + 0{,}101 \cdot Nota\_CalculoII +$$
$$+ 0{,}085 \cdot Nota\_Teoria.Maq.y.Mecanismos$$

Optimización y Simulación

$$Nota\_Optimiz.ySimulacion = -1{,}753 + 0{,}135 \cdot Nota\_Din.Sistemas + 0{,}237 \cdot Nota\_Ter\,\text{mod}\,inamica +$$
$$+ 0{,}207 \cdot Nota\_Ter\,\text{mod}\,inamica.Fund. + 0{,}253 \cdot Nota\_Tec.Estad.Calidad + 0{,}161 \cdot Nota\_CalculoI +$$
$$+ 0{,}113 \cdot Nota\_Electrotecnia + 0{,}106 \cdot Nota\_Informatica$$

Gestión de Proyectos

$$Nota\_Gestion\,\text{Pr}\,oyectos = 4{,}715 + 0{,}136 \cdot Nota\_Materiales + 0{,}101 \cdot Nota\_Tec.Estad.Calidad +$$
$$+ 0{,}065 \cdot Nota\_Optimiz.y.Simulacion + 0{,}072 \cdot Nota\_Econom.yEmpresa$$

Electrónica

$$Nota\_Electronica = 0{,}674 + 0{,}128 \cdot Nota\_Optimiz.y.Simulacion +$$
$$+ 0{,}134 \cdot Nota\_Mec.Medios.Cont. + 0{,}111 \cdot Nota\_Mecanica.Fund. +$$
$$+ 0{,}184 \cdot Nota\_Maq.Electricas + 0{,}113 \cdot Nota\_Mecanica +$$
$$+ 0{,}132 \cdot Nota\_Tec.Selec.Materiales + 0{,}129 \cdot Nota\_A\lg.Lineal$$

Sistemas de Fabricación

$$Nota\_Sist.Fabricacion = 3{,}093 + 0{,}236 \cdot Nota\_\mathrm{Re}sist.Materiales +$$
$$+ 0{,}151 \cdot Nota\_Din.Sistemas + 0{,}204 \cdot Nota\_Org.y.Gestion + 0{,}155 \cdot Nota\_CalculoI +$$
$$0{,}105 \cdot Nota\_Ter\,\mathrm{mod}\,inamica.Fund. - 0{,}141 \cdot Nota\_Electromagnetismo +$$
$$+ 0{,}136 \cdot Nota\_Ter\,\mathrm{mod}\,inamica. - 0{,}154 \cdot Nota\_\mathrm{Pr}oyectoII.$$

Termotecnia

$$Nota\_Termotecnia = -1{,}861 + 0{,}32 \cdot Nota\_\mathrm{Re}sist.Materiales + 0{,}15 \cdot Nota\_Din.Sistemas +$$
$$+ 0{,}236 \cdot Nota\_Org.y.Gestion + 0{,}206 \cdot Nota\_Maq.Electricas + 0{,}098 \cdot Nota\_Mecanica +$$
$$+ 0{,}195 \cdot Nota\_Tec.Selec.Materiales + 0{,}171 \cdot Nota\_Ec.Diferenciales +$$
$$(-0{,}102) \cdot Nota\_Fund.Informatica + 0{,}134 \cdot Nota\_Optimiz.y.Simulacion$$

Control Automático

$$Nota\_ControlAutomatico = -1{,}021 + 0{,}302 \cdot Nota\_\mathrm{Re}sist.Materiales +$$
$$+ 0{,}209 \cdot Nota\_Electomagnetismo + 0{,}21 \cdot Nota\_Tec.MedioAmb.y.Sost. +$$
$$+ 0{,}136 \cdot Nota\_Mecanica + 0{,}2 \cdot Nota\_Tec.Selec.Materiales +$$
$$+ 0{,}168 \cdot Nota\_Mec.Medios.Continuos - 0{,}107 \cdot Nota\_Exp.Grafica$$

TFG

$$Nota\_Termotecnia = 7{,}361 + 0{,}093 \cdot Nota\_Fund.Informatica + 0{,}243 \cdot Nota\_Control.Aut. +$$
$$(-0{,}18) \cdot Nota\_Optimiz.y.Simulacion + 0{,}182 \cdot Nota\_Informatica - 0{,}128 \cdot Nota\_QuimicaII$$

6.8.2. Resultados obtenidos con la aplicación de los modelos

A continuación, se muestra una tabla resumen con los resultados obtenidos aplicando los modelos mostrados en el apartado anterior:

Asignatura	% Acierto Cualificación	%Acierto Aprobado o Suspendido	Datos Totales	Error Típico
Geometría	58,10 %	74,30 %	284	1,37
Cálculo II	57,19 %	71,23 %	285	1,52
Termodinámica Fundamental	58,42 %	71,33 %	279	1,53
Química II	53,44 %	78,95 %	247	1,55
Expresión Gráfica	56,01 %	73,20 %	291	1,82
Electromagnetismo	59,47 %	70,00 %	190	1,45
Métodos Numéricos	62,57 %	87,17 %	187	1,42
Materiales	64,41 %	75,71 %	177	1,18
Ecuaciones Diferenciales	55,14 %	71,35 %	185	1,33
Informática	56,59 %	83,52 %	182	1,61
Mecánica	56,77 %	60,94 %	192	1,67
Economía y Empresa	66,00 %	94,00 %	150	0,97
Estadística	64,86 %	88,51 %	148	1,10
Dinámica de Sistemas	62,33 %	84,93 %	146	1,38
Proyecto I	54,79 %	100,00 %	146	1,00
Teoría Maquinas y Mecanismos	58,22 %	78,08 %	146	1,41
Tec. Medio Ambiente y Sosten.	57,98 %	74,79 %	119	1,18
Termodinámica	56,30 %	80,67 %	119	1,17
Electrotecnia	55,56 %	66,67 %	126	1,32
Mecánica Medios Continuos	53,85 %	72,65 %	117	1,21
Técnicas Estadísticas Calidad	68,07 %	95,80 %	119	0,91
Tec. Selección Materiales	63,33 %	78,33 %	120	1,06
Mecánica de Fluidos	65,06 %	72,29 %	83	1,11
Organización y Gestión	57,61 %	81,52 %	92	1,02
Resistencia de Materiales	72,73 %	81,82 %	88	0,97
Proyecto II	61,11 %	100,00 %	90	0,80
Máquinas Eléctricas	71,74 %	95,65 %	92	1,16
Optimización y Simulación	55,42 %	75,90 %	83	1,31
Gestión de Proyectos	65,22 %	98,55 %	69	0,62
Electrónica	57,97 %	91,30 %	69	1,12
Sistemas de Fabricación	60,00 %	95,38 %	65	1,01
Termotecnia	56,06 %	90,91 %	66	1,41
Control Automático	64,71 %	85,29 %	68	1,23
TFG	55,17 %	100,00 %	29	0,90

Tabla 6.92. Tabla resumen de los porcentajes de acierto

Los porcentajes de acierto, tanto para la cualificación como para si aprueban o suspenden, son mayores en las asignaturas de los dos últimos cursos. En promedio, el porcentaje de acierto de las cualificaciones en los subconjuntos de validación es de 60,1 %, mientras que el pronóstico de si el alumno aprobará o suspenderá se predice con un 82,4 % de acierto en promedio.

7. Evaluación económica

Se considera la realización de este proyecto como un trabajo de investigación. Los costes atribuibles al proyecto corresponden a recursos informáticos, material de oficina, y principalmente, a costes del personal. En este capítulo se realizan los cálculos necesarios para determinar un precio final del proyecto.

7.1. Costes de personal

A nivel de costes de personal, se deben diferenciar tres tareas: una de investigación, otra de ingeniería, y una última de administración. Durante la etapa de investigación la principal tarea es la de adentrarse en la materia del proyecto, recopilando la información necesaria y consultando las fuentes adecuadas. Posteriormente se pasa a la fase de ingeniería, donde se plantea el problema y se analizan las alternativas para resolverlo, seleccionando finalmente la solución óptima y contrastando los resultados. Por último, es necesario un trabajo de administración, en el que se lleva a cabo la elaboración de la memoria y la exposición de conclusiones y resultados.

En la siguiente tabla se detallan los costes de personal en función de las horas dedicadas a cada tipo de trabajo y el coste por hora.

Concepto	Coste (€/h)	Horas (h)	Coste (€)
Investigación	30	100	3000
Ingeniería	45	300	13500
Administración	20	100	2000
Total		**500**	**18500**

Tabla 7.1. Desglose de los costes de personal

7.2. Recursos informáticos

En los recursos informáticos se incluyen todos aquellos costes relacionados con la amortización del ordenador y con la obtención del software necesario. Para la realización del proyecto se ha requerido un ordenador, la licencia de software SPSS.

Se ha empleado un portátil valorado en 700 € y se considera un coste de mantenimiento del 10% anual de su precio de compra para un uso de 1200 horas anuales.

Si se establece que durante la realización del proyecto se ha usado el ordenador el 95 % del tiempo, el coste de mantenimiento queda:

$$500\ horas \cdot 0{,}95 \cdot \frac{700\ euros \cdot 0{,}1}{1200\ horas} = 27{,}71\ euros$$

Para el cálculo de la amortización, se establece que el ordenador se utiliza 48 semanas al año (52 que tiene un año menos 4 de vacaciones que debe tener el empleado) durante un periodo de 4 años. El ordenador ha sido empleado durante 11 meses, es decir, 44 semanas. Si se descuenta un día de descanso de cada semana, el uso real es de 38 semanas.

$$28\ semanas \cdot \frac{\frac{700\ euros}{4\ años}}{38\ semanas} = 128{,}95\ euros$$

El coste total asociado al ordenador es de:

Coste Ordenador = 27,71 + 128,95 = 156,66 €

Para estimar el coste de la licencia de software SPSS, lo correcto sería calcular el número de horas que se ha utilizado el programa para el proyecto y calcular el coste atribuible, ya que dicho programa se podría emplear para otros proyectos paralelos y no habría que pagar una nueva licencia. No obstante, bajo la hipótesis de que sólo se está realizando el presente proyecto, el coste de la licencia anual es de 200 €.

Por lo tanto, el coste total de los recursos informáticos es el siguiente:

Costes informáticos totales = 156,66 + 200 = 356,66 €

7.3. Material de oficina

En este apartado se incluyen los gastos relativos a fotocopias, carpetas, CDROM, etiquetas, encuadernaciones, gastos de impresión, etc. A continuación se muestra el detalle de los costes.

Concepto	Unidades	Coste Unitario (€)	Coste (€)
Fotocopias	310	0,3	93,0
Encuadernación	3	3	9,0
CD - ROM	5	0,6	3,0
Caja Proyecto	2	4	8,0
Total			**113,0**

Tabla 7.2. Desglose costes de oficina

7.4. Coste total del proyecto

El coste total del proyecto se constituye de la suma de los distintos costes calculados en los apartados anteriores. Se muestra en la siguiente tabla:

Concepto	Coste (€)
Costes de personal	18.500
Recursos informáticos	356,66
Material de oficina	113,00
Total	**18.969,66**

Tabla 7.3. Coste total del proyecto

8. Impacto medioambiental

La realización de este proyecto no requiere de un análisis exhaustivo del impacto medioambiental, ya que se ha realizado íntegramente con un ordenador y no se ha generado ningún tipo de residuo.

El único impacto ambiental a tener en cuenta es el producido durante la ejecución del proyecto, ya que se requiere de electricidad para alimentar el ordenador y el router para la conexión a internet. También se debe considerar el gasto energético para alumbrar el puesto de trabajo, pero es insignificante si se considera que se ha aprovechado las horas de luz natural y que muy probablemente, el usuario necesita la luz aunque no esté trabajando.

Por último, remarcar que no se debe contemplar en el impacto medioambiental el gasto de papel producido por la impresión de la memoria y de los anexos, ya que dicho coste es contemplado en la evaluación económica del proyecto.

9. Conclusiones

El objetivo del presente proyecto era realizar un análisis en profundidad de los resultados obtenidos por los estudiantes de l'Escola Tècnica Superior d'Enginyeria Industrial de Barcelona (ETSEIB). La finalidad última de dicho análisis es la construcción de distintos modelos que permitan determinar el comportamiento del alumno en un futuro. Con la palabra comportamiento, se hace referencia a una serie de aspectos tales como si el alumno repetirá o no en los distintos cursos, si estará por encima o por debajo de nota media de cada curso, o si aprobará los cursos al ritmo estipulado (curso por año). Como fruto de dicho análisis, también se han construido distintas ecuaciones con el fin de estimar la nota que obtendrá el alumno en cada una de las asignaturas .

Los datos de origen para el proyecto pertenecen a las tres titulaciones de grado impartidas en la ETSEIB, que son Grado en Ingeniería en Tecnologías Industriales, Grado en Ingeniería Química y Grado en Ingeniería de Materiales. No obstante, el estudio en profundidad de este proyecto se ha centrado únicamente en la primera de ellas, debido a que para las otras dos titulaciones no se dispone de una cantidad de datos suficiente para obtener unas conclusiones concluyentes.

El primer paso necesario para realizar el análisis consiste en transformar la base de datos original. Para ello, se han procesado los datos, excluyendo aquellos casos que puedan distorsionar los resultados, creando nuevas variables, y disponiendo los casos útiles de manera que se facilite su análisis. Finalmente, para el análisis de la titulación Grado en Ingeniería en Tecnologías Industriales se dispone de 54.993 registro correspondientes a un universo de 1820 alumnos, de los cuales un 77,64 % son hombres y un 22,36 % son mujeres.

Los promedios de los distintos cursos se mantienen entre 6,30 y 6,60, salvo para el cuarto y último curso, que el promedio es de 7,54. Se observa también un claro descenso del promedio de asignaturas repetidas en cada curso, que va desde 2,49 en primero a 0,12 en cuarto. Además de por la fase selectiva, uno de los motivos de este decrecimiento es el filtro natural sobre los alumnos, por el cual aquellos con peores resultados abandonan los estudios.

La nota promedio más alta de la titulación se encuentra en el Trabajo de Fin de Grado (TFG) con un 8,73. Le sigue de cerca la asignatura del tercer curso Proyecto II, con una nota media de 8,42. En ninguna de las dos asignaturas se ha registrado un suspenso en ninguno de los años que se poseen datos. Por el contrario, la asignatura con mayor porcentaje de suspensos de la titulación es la asignatura de segundo Mecánica, con un 48,74 % de

suspensos en las convocatorias registradas. Además, la nota media más baja también coincide con dicha asignatura con un valor de 4,33.

Concluido el análisis de tipo descriptivo, se procede al estudio en detalle de las primeras variables, que son tres: el sexo, el lugar de residencia del alumno, y el lugar de la escuela donde el alumno ha cursado sus estudios. El objetivo es doble: observar si existen diferencias significativas en los resultados entre los grupos de las variables (hombre y mujer para el sexo, y distintas provincias para las ubicaciones) y determinar si dicha variable es buena predictora para los modelos. Las técnicas empleadas son dos: la U de Mann Whitney para dos muestras independientes para el sexo, y el test de Kruskal Wallis para k muestras independientes para las ubicaciones. En cuanto al sexo, no hay diferencias significativas entre hombres y mujeres, por lo que no es una buena variable predictora. En cuanto al lugar de residencia del alumno y el emplazamiento de la escuela donde ha estudiado, se encuentran diferencias significativas en el número de asignaturas repetidas en primero y en el parámetro Alpha de primero. Además, en el caso del lugar de residencia del alumno, también se observan diferencias significativas en las notas promedio de primero. Por lo tanto, ambas variables son consideradas para construir el modelo para determinar el comportamiento en primero.

La técnica empleada para la construcción de los modelos que permitan clasificar a los estudiantes es la regresión logística binaria. Dicha técnica constituye una alternativa a la regresión lineal cuando alguna de las variables tratadas son categóricas. Se han contemplado otros métodos de clasificación como por ejemplo el análisis discriminante, pero se ha descartado su uso por ofrecer resultados menos precisos que la regresión logística binaria. Las variables empleadas para la construcción del modelo varían en función del curso a pronosticar. En el primer curso, las únicas variables disponibles son la nota de la selectividad, el lugar de residencia del alumno, y el lugar de la escuela donde el alumno ha estudiado. En cambio, para cursos posteriores, se emplean otras variables relativas a los resultados obtenidos por el alumno en cursos anteriores, como por ejemplo el parámetro Alpha, el promedio obtenido en el curso anterior, el número de asignaturas repetidas, etc.

Cada uno de los modelos generados ha sido validado con un subconjunto de datos que no ha participado en la creación del mismo para garantizar su precisión. Los porcentajes de acierto obtenidos son elevados, siendo mayores para cursos posteriores a primero debido al mayor número de variables predictoras. En cuanto a la predicción de estudiantes que aprobarán los cursos al ritmo estipulado (curso por año), se obtienen unos porcentajes de acierto del 70,8 % en el primero curso, y de alrededor del 86 % para cursos posteriores. En cuanto al pronóstico de si un alumno repetirá o no en los distintos cursos, los alumnos clasificados con éxito están alrededor de 72 % en primero, y alrededor del 80 % en los cursos siguientes. Por último, el porcentaje de acierto de los alumnos que estarán por

encima o por debajo de la media en los distintos cursos está alrededor del 70 % en el primer y segundo curso, y alrededor del 80 % en tercer y cuarto curso.

Para realizar las estimaciones de las notas obtenidas por el estudiante en cada una de las asignaturas se ha empleado la técnica estadística de la regresión lineal. Se ha utilizado dicha técnica debido a que todas las variables empleadas son cuantitativas, y a que todas ellas están en una misma escala (de 0,0 a 10,0). Para cada una de las asignaturas, se calcula la nota que obtendrá el alumno mediante una combinación lineal de las notas obtenidas en las asignaturas de cursos anteriores. Los resultados varían en función de la asignatura pronosticada, pero en términos globales se obtiene un error típico de 1,23. Con el objetivo de determinar si los casos de notas extremas perturbaban mucho los resultados, se han calculado los porcentajes de acierto obtenidos codificando las notas observadas y predichas en suspensos, aprobados, notables, y sobresalientes. El porcentaje de acierto promedio que se ha obtenido es del 60,1 %. Por último, se ha repetido el proceso pero con una clasificación todavía más genérica: aprobados y suspendidos. El porcentaje de acierto promedio es de 82,4 %.

Conocer qué resultados obtendrán los alumnos en un futuro representa una gran ventaja. Por ejemplo, se podría ofrecer un refuerzo a aquellos estudiantes que se predice que tendrán más dificultades, o se podría subdividir la clase en grupos con el objetivo de aumentar la tasa de éxito y así aumentar el prestigio de la universidad. Conocer qué asignaturas suspenderá también es de gran utilidad, ya que permite realizar un análisis de flujos y adaptar los recursos de espacio y de personal de una manera más efectiva.

10. Bibliografía

Bibliografía complementaria

Regresión Logística con SPSS “paso a paso”. Aguayo Canela, Mariano.

(http://www.fabis.org/html/archivos/docuweb/Regres_log_1r.pdf)

Video tutorial Regresión Logística.

(https://www.youtube.com/watch?v=foF2-6wLbYk)

Video tutorial Regresión Logística con SPSS

(https://www.youtube.com/watch?v=iOBPDEFZMLM)

Tutorial de Regresiones de IBM

(http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/SPSS19/IBM-SPSS_regression.pdf)

Estadística no paramétrica: Prueba de Chi - Cuadrado. Juan Francisco Monge Ivars, Ángel A. Juan Pérez

(http://www.uoc.edu/in3/emath/docs/Chi_cuadrado.pdf)

Prueba de Chi - Cuadrado

(http://es.slideshare.net/armando310388/prueba-chicuadrado)

Prueba de la U de Mann - Whitney. Wikipedia.

(https://es.wikipedia.org/wiki/Prueba_U_de_Mann-Whitney)

Pruebas para dos muestras independientes. Universidad de Barcelona.

(http://www.ub.edu/aplica_infor/spss/cap6-2.htm)

La prueba U de Mann Whitney. Maribel Correa Taborda y Sandra P. Vallejo Florez.

(http://slideplayer.es/slide/1698456/)

Pruebas para K muestras independientes. Universidad de Barcelona.

(http://www.ub.edu/aplica_infor/spss/cap6-4.htm)

Prueba no paramétrica de Kruskal Wallis. Irene Soler, Abel W. Reyes, Víctor M. García.

(http://www.estadisticafi.unam.mx/point/10.pdf)

Regresión Logística Binaria. Ángel Alejandro, Juan Pérez, Renatas Kizys, Luis María Manzanedo Del Hoyo.

(http://www.uoc.edu/in3/emath/docs/T10_Reg_Logistica.pdf)

Apuntes de Bioestadística. F. J. Barón López, F. Téllez Montiel.

(http://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap08.pdf)

Medidas de Distribución - Asimetría y Curtosis

(http://www.spssfree.com/curso-de-spss/analisis-descriptivo/medidas-de-distribucion-curtosis-asimetria.html)

Regresión lineal con SPSS. Escuela Superior de Informática.

(https://www.uclm.es/profesorado/raulmmartin/Estadistica/PracticasSPSS/REGRESION_LINEAL_CON_SPSS.pdf)

Minería de datos. Wikipedia.

(https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos)

Modelo de Regresión lineal múltiple. Renatas Kizys, Ángel A. Juan.

(http://www.uoc.edu/in3/emath/docs/T01_Reg_Lineal_Multiple.pdf)

Regresión lineal múltiple. J. M. Rojo Abuín.

(http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion_lineal_multiple_3.pdf)

El Análisis de la Regresión a través de SPSS. M.Dolores Martínez Miranda

(http://www.ugr.es/~curspss/archivos/Regresion/TeoriaRegresionSPSS.pdf)

La media Aritmética. Vitutor.

(http://www.vitutor.com/estadistica/descriptiva/a_10.html)

La varianza. Vitutor.

(http://www.vitutor.com/estadistica/descriptiva/a_15.html)

Regresión Logística. Santiago de la Fuente Fernández.

(http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/CUALITATIVAS/LOGISTICA/regresion-logistica.pdf)